EUROPEAN LANGUAGE EQUALITY 2

FSTP Project Report

GL-BLARK – A BLARK for minoritized languages in the era of deep learning: expertise from academia and industry

Authors	Sofía García[1] and Iria de-Dios-Flores[2]
Organisation	[1]Factoría de Software e Multimedia, S. L. (imaxin|software), [2]Universidade de Santiago de Compostela, Centro Singular de Investigación en Tecnoloxías Intelixentes

Dissemination level	Public
Date	03-04-2023

About this document

Project	European Language Equality 2 (ELE2)
Grant agreement no.	LC-01884166 – 101075356 ELE2
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-07-2022, 12 months
FSTP Project	GL-BLARK – A BLARK for minoritized languages in the era of deep learning: expertise from academia and industry
Authors	Sofía García[1] and Iria de-Dios-Flores[2]
Organisation	[1]Factoría de Software e Multimedia, S.L. (imaxin|software), [2]Universidade de Santiago de Compostela, Centro Singular de Investigación en Tecnoloxías Intelixentes
Type	Report
Number of pages	44
Status and version	Final
Dissemination level	Public
Date of delivery	03-04-2023
EC project officer	Susan Fraser
Contact	European Language Equality 2 (ELE2) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality 2 (ELE2) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
5	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
6	European Federation of National Institutes for Language	EFNIL	LU
7	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR

Contents

1. Introduction **1**

2. The concept of BLARK: previous works **2**

3. Objectives and Methodology **4**
3.1. Objectives 4
3.2. Methodology 4

4. The present BLARK **6**
4.1. Cross-cutting resources 10
4.1.1. Corpora 10
4.1.2. Lexical resources 11
4.1.3. NLP tools 13
4.1.4. Language Models 13
4.2. LT tasks 16
4.2.1. Speech synthesis 16
4.2.2. Speech recognition 17
4.2.3. Machine translation 18
4.2.4. Other LT tasks 20
4.2.4.1. Grammatical error correction 20
4.2.4.2. Summarization 21
4.2.4.3. Sentiment Analysis 22
4.2.4.4. Fact checking 23
4.2.4.5. Dialog systems 24
4.2.5. Benchmarking 25

5. GL-BLARK **26**

6. Conclusions and Limitations **28**

A. Appendix: The BLARK **32**

B. Appendix: The BLARK for Galician (GL-BLARK) **36**

List of Figures

List of Tables

1. BLARK matrix structure and weights (expressed in percentages). 6
2. Cross-cutting resources: corpora. Note that corpora are worth 70% of the cross-cutting resources category. 7
3. Cross-cutting resources: corpora. Note that corpora are worth 70% of the cross-cutting resources category. 10
4. Cross-cutting resources: lexical resources. Note that lexical resources are worth 10% of the cross-cutting resources category. 12
5. Cross-cutting resources: NLP tools. Note that NLP tools are worth 10% of the cross-cutting resources category. 13
6. Cross-cutting resources: language models. Note that language models are worth 10% of the cross-cutting resources category. 14
7. LT tasks: speech synthesis. Note that speech synthesis is worth 20% of the LT tasks category. 17
8. LT tasks: speech recognition. Note that speech recognition is worth 20% of the LT tasks category. 18
9. LT tasks: machine translation. Note that machine translation is worth 20% of the LT tasks category. 19
10. LT tasks/Other LT tasks: grammatical error correction. Note that grammatical error correction is worth 20% of the Other LT tasks subcategory. 20
11. Summarization table evaluated according to corpora and models. LT tasks/Other LT tasks: summarization. Note that summarization is worth 20% of the Other LT tasks subcategory. 21
12. LT tasks/Other LT tasks: sentiment analysis. Note that sentiment analysis is worth 20% of the Other LT tasks subcategory. 22
13. LT tasks/Other LT tasks: fact checking. Note that fact checking is worth 20% of the Other LT tasks subcategory. 23
14. LT tasks/Other LT tasks: dialog systems. Note that systems is worth 20% of the Other LT tasks subcategory. 24
15. LT tasks: benchmarking. Note that benchmarking is worth 10% of the LT tasks category. 26

List of Acronyms

AGPL	Affero General Public License
ASR	Automatic Speech Recognition
BERT	Bidirectional Encoder Representations from Transformers
BLARK	Basic Language Resource Kit
BLEU	Bilingual Evaluation Understudy
CiTIUS	*Centro Singular de Investigación en Tecnoloxías Intelixentes*
CORGA	Current Galego Reference Corpus
CREA	Reference Corpus of the Current Spanish
CTG	Galician Technical Corpus
CTILC	Computerized textual corpus of the Catalan language
DLDP	Digital Language Diversity Project

DLE	Digital Language Equality
ELDA	Evaluations and Language resources Distribution Agency
ELE	European Language Equality
ELE2	European Language Equality *(this project)*
ELEN	European Language Equality Network
ELG	European Language Grid (EU project, 2019-2022)
ELRA	European Language Resource Association
ELSNET	European Network of Excellence in Language and Speech
GEC	Grammatical Error Correction
GLEU	Generalized Language Understanding Evaluation
GPL	General Public License
GPT	Generative Pre-trained Transformer
ILG	*Instituto da Lingua Galega*
LT	Language Technology/Technologies
M2M	Many-to-Many
MCR	Multilingual Central Repository
ML	Machine Learning
MOS	Mean Opinion Score
MT	Machine Translation
MTP	Münster Tagging Project
NEMLAR	Network for Euro-Mediterranean LAnguage Resource
NERC	Name Entity Recognition and Classification
NLLB	No Language Left Behind
NLP	Natural Language Processing
NMT	Neural Machine Translation
OSCAR	Open Super-large Crawled ALMAnaCH coRpus
POS	Part-Of-Speech
RAG	Dictionary of the Royal Galician Academy
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
TTS	Text To Speech
WER	Word Error Rate

Abstract

This desk research addresses the complex task of defining what could be an appropriate **B**asic **LA**nguage **R**esource **K**it (BLARK) for minoritized languages in the deep learning era. To do so, we take Galician as a starting point being a paradigmatic case of a minoritized language in Europe, with the ultimate goal of generating a BLARK that is language independent. We define the minimum linguistic resources and tools in order to develop LTs in the field of language-centric artificial intelligence in order to establish a starting point that helps identify which key areas would need to be developed in order to cover the basic LTs, define areas of potential growth, and help funding bodies spot investment needs. The BLARK is divided in two broad categories: cross-cutting resources and LT tasks. The former covers those fundamental resources of a more general nature, while the latter covers the resources needed for specific tasks or applications. Besides the BLARK matrices, an additional contribution of the project is the release of the first BLARK for Galician (GL-BLARK), with the further goal that it serves as a reference source for other minority languages in the European context.

1. Introduction

This desk research addresses the complex task of defining what could be an appropriate **B**asic **LA**nguage **R**esource **K**it (BLARK) for minoritized[1] languages in the deep learning era. In recent years, the explosion of deep learning methods has brought about a huge paradigm shift in the area of Language Technologies (LTs), radically changing the field, and most importantly, moving LTs to a state of development that was hard to foresee just a few years earlier. Nonetheless, these advancements have not been comparable across languages, generating a gap that is being aggravated by the need for the large quantities of data that modern deep learning techniques require. With this in mind, this BLARK proposal focuses on minoritized languages, taking Galician as a starting point being a paradigmatic case of a minoritized language in Europe, with the ultimate goal of generating a BLARK that is language independent.

This BLARK aims at defining the **minimum requirements** in terms of linguistic resources and tools in order to develop LTs in the field of language-centric artificial intelligence. The purpose is to establish a starting point that helps identify which key areas would need to be developed in order to cover the basic LTs, define areas of potential growth, and help funding bodies spot investment needs. For these reasons, this BLARK is not an appropriate tool to measure the degree of development of high-resourced languages such as English or Chinese, or even languages that have a moderate amount of support such as German, Spanish, or Portuguese, where most of the minimum technological developments are covered, even when they still need significant support to keep growing. Furthermore, considering the speed at which the field of LTs is evolving nowadays, this proposal faces the challenge of becoming obsolete soon, as new solutions are constantly emerging. Consequently, this BLARK will need constant revision and updating. Another important aspect that determined the design of the BLARK proposed has to do with its usability. In this respect, our intention was to create a BLARK whose completion would not be excessively time-consuming or would require a great amount of inquiry and research. The objective in this regard is that anyone in the field of LTs with expert knowledge about a given language would be able to fill it in a timely manner. To achieve this, we have condensed or simplified some aspects that, as we know, are far from simple but that, on the other side, would allow for more flexibility to adapt to new realities.

[1] We have favored the use of the term 'minoritized' instead of 'minority' because, as mentioned by Suay and Hicks (2018), this term provides a better characterization of the situation of several languages in Europe (e.g. Catalan or Galician) which have a sizable number of speakers in their own territories and hence, are not a 'minority', even if they are not dominant languages.

On top of the above, we intended to create a BLARK that can cover the key resources from the perspective of academia and industry. At a time when the most important factor for developing new technologies is data, having access to data, especially in minoritzed languages, is radically different whether the goal is to use it for research or for product development. The complementary perspectives of industry and academia are combined thanks to the collaboration between **imaxin**|*software*, a leading company in the language technology sector in Galicia since 1997, and the scientific advisory of the Nós project (*Proxecto Nós*)[2], a public initiative by *Xunta de Galicia* and the *Universidade de Compostela* aimed at providing the Galician language with openly licensed resources and tools in the area of intelligent technologies (see de Dios-Flores et al. (2022)). Besides the BLARK matrices themselves, an additional contribution of the project is the release of the first BLARK for Galician (GL-BLARK), with the further goal that it serves as a reference source for other minority languages in the European context. As stated in the report D1.15 of the European Language Equality Project (Sánchez and Mateo, 2022), Galician is currently the language with the least technological support in Spain and one of the least supported in Europe. Thanks to the efforts of the Nós Project, the situation has improved from the publication of Sánchez and Mateo (2022) report, as will be shown in the GL-BLARK, although the outlook continues to be that of a language in much need of support.

The structure of this report is as follows: in Section 2 we provide a review of the concept of BLARK and revise previous BLARKs. In Section 3, we describe our objectives (3.1) and the methodology deployed to design this BLARK proposal (3.2). Then, the central contribution of this project is found in Section 4, where we explain in detail the structure and categories of the proposed BLARK, and justify the importance given to each of the resources covered and their relevant quality aspects. In order to test the applicability of this BLARK, we put Galician to the test, and report the results in Section 5. Finally, in Section 6, we present some conclusions and analyze the limitations of this work. The BLARK matrices as well as a filled version for Galician (GL-BLARK) are to be found, respectively, in appendices A and B.

2. The concept of BLARK: previous works

In 1998 ELSNET[3] and ELRA[4] presented the work titled *ELSNET and ELRA: Common past, common future* (Krauwer, 1998), where the concept of BLARK was initially proposed as the first step to measure the development of LT resources. This first attempt was mostly envisaged as a means to cover the reality of major languages from Central and Eastern Europe countries, although the idea was to make it useful to any lesser favored languages. The key language-independent resources covered by this initial BLARK proposal and definition were the following (Krauwer, 1998, p.3):

> a) The minimum general text corpus required to be able to do any precompetitive research.
>
> b) Something similar for a spoken text corpus.
>
> c) A collection of basic tools to manipulate and analyze the corpora
>
> d) A collection of skills that constitute the minimal starting point for the development of industry

[2] The execution of *Proxecto Nós* is currently being carried out by a research team comprising members of the *Instituto da Lingua Galega* (ILG) and the *Centro Singular de Investigacion en Tecnoloxías Intelixentes* (CiTIUS).

[3] European Network of Excellence in Language and Speech, founded in 1991 as a pilot Network by the European Commission's ESPRIT Long Term Research programme to construct multilingual integrated language and speech systems with unrestricted coverage of spoken and written language .

[4] The European Land Registry Association was created in 2004 by a group of Land Registry organizations. ELRA wants to promote a mutual understanding of land registers, to help create an open and secure Europe, serving and protecting citizens.

Upon the definition of the BLARK, the proposed first step was to identify what were those LT resources that already existed for each language and, once all this information was collected, the next step would be to launch different projects to fill the gaps identified in the BLARK for each of those languages (Krauwer, 1998). Hence, the BLARK was conceived as a starting point to systematically identify the gaps. Four years later, in 2002, the Dutch Language Union (Nederlandse Taalunie-NTU) launched the first BLARK in Dutch (Binnenpoorte et al., 2002a). Following the steps defined by Krauwer (1998), they described the basic LT resources for Dutch that should be available both for academia and industry. Binnenpoorte and colleagues divided their Dutch BLARK in three main categories: applications (the services that make use of LTs), modules (the basic software components essential for developing LT applications), and data (the data needed to build the modules). Taking these three key areas as a starting point, they generated a BLARK matrix that covered the modules needed for the applications, the data needed for the modules, and the relative importance of modules and data. Some of the conclusions reached by Binnenpoorte et al. (2002a, p.1864) in this first BLARK were that, for text applications, the most important modules were those for preprocessing the text (e.g. tokenizetion, name entity recognition, syntactic and semantic analysis, etc.). With respect to data, the most important resources were mono-lingual lexicons, annotated corpora, and benchmarks for evaluation. Regarding speech, the modules for automatic speech recognition, speech synthesis, and identification of confidence measures had the most relevance. Finally, in speech data, they identified multi-modal, multi-media and multi-lingual speech corpora, in addition to speech corpora for specific applications and benchmarking. They used this BLARK matrix for a survey of the Dutch and Flemish languages. Some of the conclusions of that survey were that more cooperation between universities, research institutes, and companies was necessary because the maintenance of the BLARK was something essential and the LTs structure was scattered, incomplete, and not sufficiently accessible (Binnenpoorte et al., 2002b).

Another well-known BLARK is the one launched by NEMLAR for Arabic languages in 2006 (Maegaard et al., 2006), whose results are also available on ELDA. They took the Dutch BLARK as a starting point and divided the BLARK between BLARK definition (general concepts) and BLARK specification (instantiation for a given language, in this case, Arabic). They also added some improvements to the BLARK that were not covered in Binnenpoorte et al. (2002a). Among the most relevant ones, they included challenges related to language standards and measures of quantity and quality for modules and data - which they determined by revising the available corpora for Arabic languages and defining the desirable size. Another difference with previous BLARKs is that Maegaard et al. (2006) created different matrices for written and spoken language, although they also related the modules with the possible applications, as Binnenpoorte et al. (2002a) did.

In 2019, as part of the Faroese ASR project, Simonsen et al. (2022) developed a BLARK for Faroese. This project finished in the summer of 2022 and all the resources are publicly available on the OpenSLR webpage. In their work, they list all the available resources for Faroese, highlighting those created under the auspices of their project. They created the EvalBlark tool, which takes as input typical BLARK resources (corpora, tools, etc.) and returns a list of reporting inconsistencies with possible corrections. Rather than being a conceptual resource that allows evaluating the support and needs of a given language, EvalBlark provides some criteria (e.g. interoperability, formatting, etc.) that the resources to be included in a BLARK should fulfill.

Despite the interest and usefulness of a tool like BLARK as a means to analyze and compare LT development across languages, efforts to develop systematic, wide-covering and language-independent BLARKs have been scarce. The existing BLARKs have only been deployed in the context of a few languages (at least publicly), and what is most relevant, they are not fully appropriate to cover LT needs in the era of deep learning. From the first approaches to develop a BLARK for minoritized languages two decades ago, the field of language-centric

artificial intelligence has experienced a complete paradigm shift that made the previous BLARK matrices and definitions somehow obsolete, especially from the explosion of neural networks applied to LTs. In our view, these new methods require different perspective-taking, making it less useful to divide a BLARK into applications, modules, and data, nor focus on its interdependence as done in previous works (Binnenpoorte et al., 2002a), (Maegaard et al., 2006). In the deep learning era, data availability has reached even greater importance, and it is for this reason that it is the cornerstone of the present BLARK.

3. Objectives and Methodology

In this section, we will describe the objectives of the project (3.1) and the methodology used to create this new BLARK (3.2). The contents of the BLARK and a detailed explanation of the selection criteria and the instructions to fill it in are detailed in Section 4.

3.1. Objectives

As was mentioned in the introduction, the main objective of our project is to update the BLARK to the deep learning era, focusing on the basic requirements that any minoritized language needs to achieve to provide its speakers with basic LT services. With this BLARK we aim to achieve four main goals:

1. Analyze the key aspects in terms of LTs resources (data, models, etc.) for minoritized languages to prosper in the digital world both from the perspective of the state-of-the-art in research and the industry needs in order to deploy end-user products.
2. Create a BLARK matrix that covers the basic resources, tools, tasks, and quality criteria that under-resourced languages can use as a reference to evaluate their current state and trace their road map toward digital language equality.
3. Provide the LT community with a BLARK that is wide-covering and flexible, while at the same time, straightforward, easy to complete, and to update.
4. Create a digital version of the BLARK matrix using the format of a web-based questionnaire, where anyone could fill in the BLARK information about any given language and obtain a BLARK matrix with the scores as output. This would facilitate the collection and management of the information, and enable comparative studies. This digital tool will be created upon having received critical feedback on this BLARK proposal.

The ultimate purpose is that, upon the completion of the BLARK, any LT actor would be able to get a panoramic view of the state of development of LTs for a given language and identify the fields where further efforts and investments are needed. Keeping the BLARK up-to-date is a critical factor that would determine its usefulness, something which is a great challenge at a time when the state of the art is constantly changing and evolving. To facilitate this, we tried to find the balance between not being too precise, so that the information provided is not quickly outdated, and, at the same time, collecting all the necessary information in order to approximate the degree of development of any given minoritized language.

3.2. Methodology

By taking other existing BLARKs as a starting point, such as those for Dutch (Binnenpoorte et al., 2002a) or Faroese (Simonsen et al., 2022), we have developed an analytical method in order to determine the basic requirements and tools for any language to get closer to the

state of the art in LTs. To do so, the methodology used had three main steps: first, we analyzed previous BLARK proposals in depth and identified which of the previously proposed criteria were still prevailing in the deep learning era and which were outdated. Second, we surveyed the available resources for different European languages in order to tear apart the minimum requirements from the desirable maximums. Third, we deployed our BLARK using our knowledge from Galician in order to test its capacity to appropriately produce a realistic and informative panorama of its degree of development.

With respect to the first step, we analyzed previous BLARKs in detail and tried to establish the axes that this new BLARK matrix should integrate. Upon great reflection, and given that the need for large amounts of data was transversal, our BLARK was divided into two great categories: cross-cutting resources, and LT tasks. This new division merges previous BLARKs' "data" and "modules" as cross-cutting resources, and keeps resources formerly referred to as "applications" as an independent category now deemed LT tasks (which cover key areas such as machine translation, speech synthesis, etc.). Many of the key aspects raised by Binnenpoorte et al. (2002a) are indeed still present, but the reality of new deep learning models calls for different ways to relate data and models, the former having much more weight. For instance, pre-trained models existing for high-resource languages (or multilingual ones) can be re-trained for low-resource languages if enough data is available, making the existence of data far more crucial than the availability of the model itself. Furthermore, following Maegaard et al. (2006), this BLARK integrates the dimensions of quality and quantity in order to evaluate the resources, and adds the type of license as a further dimension since it is a key factor for the development of research and end-user products by the industry. The different levels for the dimensions of quantity, quality, and license were determined in the second step of the method, by surveying the existing resources in other languages. For each of the score ranges, we provide a series of qualitative and quantitative definitions. It should be noted, however, that the criteria introduced in the BLARK, and described in detail in the next section, should be simply taken as a guide to orient, rather than rigid criteria.

With respect to the second step, which required surveying the available resources for different low-resource and high-resource languages in order to set a realistic reference, the ELG Catalogue and the different language-specific ELE deliverables were highly valuable assets that helped us locate existing resources for our different reference languages and categories. For instance, in order to determine the desired size and quality of a reference corpus, we compared the main reference corpora for Galician (CORGA), Spanish (CREA) or Catalan (CTILC). This information helped us establishing a size and quality range to set the small, medium and high score ranges. In order to evaluate the size in terms of parameters and training corpus for a language model, we compared the existing language models for English (e.g. BERT, (Devlin et al., 2019)), Basque (e.g. Berteus, (Agerri et al., 2020a)) or Galician (e.g. Galician BERT, (Garcia, 2021)). In this regard, this BLARK is envisaged as a complement to the ELE Catalogue and the language-specific ELE reports that already provide a detailed survey of the degree of development of dozens of languages in Europe, such that it could be taken as a systematic and quantitative tool that would facilitate a fast and easy comparison across minoritized languages.

As was already mentioned, what is a minimum requirement today may be completely useless as an indicator in a few years' time, or may depend on the specific language and research context (e.g. some models can be successfully fine-tuned with fewer data, or some low-resource languages or variants may benefit the most from transfer learning from languages that are typologically similar). Therefore, when it comes to completing the BLARK, we ultimately leave it to the judgment of the researcher to decide whether, for instance, a corpus is large in size or high in quality, regardless of the definitions provided as a guide in this report. This flexibility would also allow LT actors to deploy this BLARK in the context of different languages although, we insist, it is intended for minoritized ones and would not serve its desired purposes when applied to medium or high-resourced languages that would

require more demanding criteria.

4. The present BLARK

This central section describes the structure, definition and criteria of the proposed BLARK, and represents the key contribution of the project. This BLARK is divided into two main categories: cross-cutting resources and LT tasks (described in detail in Sections 4.1 and 4.2, respectively). The former covers those fundamental resources of a more general nature, while the latter covers the resources needed for specific tasks or applications. Even though the ability to successfully perform on the different LT tasks partly depends on the availability of high-quality cross-cutting resources, LT tasks were given greater weight because they cover a wide variety of applications that cannot be built without specific data (and models) of their own. These two big categories are divided into several subcategories, as shown in Table 1. On the one hand, cross-cutting resources are divided into corpora (4.1.1), lexical resources (4.1.2), NLP tools (4.1.3), and language models (4.1.4). On the other hand, LT tasks are divided into speech synthesis (4.2.1), speech recognition (4.2.2), machine translation (4.2.3), other LT tasks (4.2.4), and benchmarking (4.2.5). Considering that this BLARK is specifically geared towards minoritized languages, we have prioritized speech synthesis, speech recognition, and machine translation by giving them more weight, because, in our view, these are the tasks for which even minoritized languages should be developed with sufficient quality in order to facilitate basic human-computer interaction. Other LT tasks such as summarization, sentiment analysis, dialog systems, etc. are covered within the subcategory of "Other LT tasks". Although we consider them important, their development in minoritized languages would require that the language is at a good stage of development. Finally, we have included a final section devoted to resources for quality measurement (i.e. benchmarking). Although quality evaluation also applies to some of the resources covered within the cross-cutting resources category, it was included within the category of LT tasks because the need for evaluation datasets is common and essential to all the tasks included there.

BLARK MATRIX **100%**	Cross-cutting resources **40%**	Corpora **70%**
		NLP tools **10%**
		Lexical resources **10%**
		Language models **10%**
	LT tasks **60%**	Speech synthesis **20%**
		Speech recognition **20%**
		Machine translation **20%**
		Other LT tasks **30%**
		Benchmarking **10%**

Table 1: BLARK matrix structure and weights (expressed in percentages).

Each of the subcategories represented in Table 1 is made up of a series of resources (or further subcategories), which will be described and explained in detail in the following sections. Yet, before moving into the detailed description of each of the subcategories and resources, some general comments are deemed with respect to the scoring system, the concepts of size, quality, license, and what is considered a "resource" in this BLARK.

Scoring system

With respect to the scoring of the BLARK, the relevance or weight given to each of the categories, subcategories, and resources is expressed in percentages, which are calculated rela-

tive to each resource, subcategory, and category in order to produce the final BLARK score. To provide a clearer example of how the scoring is performed, Table 2 advances the structure and resources covered within the ”corpora” subcategory.

Resource	**Size 30%**	**Quality 40%**	**License 30%**
Annotated corpora **20%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
	Size 40%	**Quality 30%**	**License 30%**
Reference corpus **30%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
Macro corpus **50%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	

Table 2: Cross-cutting resources: corpora. Note that corpora are worth 70% of the cross-cutting resources category.

The specific description of the resources and evaluation criteria included for the corpora subcategory are detailed in Section 4.1.1, but what concerns us now is to provide an example of how the scoring would be performed. As shown in Table 1, corpora receive 70% of the importance within the cross-cutting resources category (which in turn is worth 40% of the final BLARK). Within the corpora subcategory, three resources are covered (i.e. annotated corpora, reference corpus, and macro corpus), each with a different associated weight (20%, 30% and 50%, respectively). Furthermore, for each resource, three aspects are evaluated in this case: size, quality, and license, each with a different associated weight (30%, 40%, and 30% in the case of annotated corpus and 40%, 30% and 30% in the case of reference and macro corpus). The difference in the quality and quantity weights reflects the fact that, in the case of annotated corpora, quality tends to be more important than size. Especially because the size needed for fine-tuning (the situations where the annotated corpora are most commonly used) is not very large compared to the size needed in a reference corpus or a macro corpus. These concepts and their associated weights will vary depending on the resource, but always add up to a total of 100% for each resource. To exemplify how this system would work, let us take the case of Galician.

Recently, a series of Galician annotated corpora with more than 4M entries have been made publically available (SLI_NERC_Galician and SLI_CTG_POS), but these only cover POS-tagging or NERC annotations. Therefore, the variety of annotations is not very wide. Consequently, Galician can be considered to have a large, low-quality and open-licensed annotated corpus. In the case of the reference corpus, CORGA, the Current Galician Reference Corpus, has 43M words. This corpus includes Galician texts from 1975 to the present day, so it can be considered a medium-sized, medium-quality corpus, but it is only open for consultation online, thus, its license would be closed. Finally, the biggest Galician macro corpus is the SLI GalWeb1.0. This corpus has around 174M words crawled from different web pages. Hence, it is a medium-size, low quality and open-licensed corpus. Consequently, regarding Galician corpora, its BLARK score would be calculated as shown below, totaling up to a 57,55% score of the corpora subcategory, which will contribute to a 40,28% within the cross-cutting resources category, and 16,11% within the global BLARK.

- **Annotated corpus** (20%): SIZE (100% (large) x 30% (size)) + QUALITY (20% (low) x 40% (quality)) + LICENSE (100% (open) x 30% (license))= 68%

- **Reference corpus** (30%): SIZE (60% (medium) x 40% (size)) + QUALITY (60% (medium) x 30%(quality)) + LICENSE (15% (closed) x 30% (license))= 46,5%
- **Macro corpus** (50%): SIZE (60% (medium) x 40% (size)) + QUALITY (20% (low) x 30%(quality)) + LICENSE (100% (open) x 30% (license))= 60%

When there is no existing resource in a given subcategory, not selecting any of the levels equates to a score of 0%. Although the scoring process seems somehow tedious, it is not intended to be hand-calculated. The digital implementation of the BLARK would facilitate the automatic generation of the scores for each resource, subcategory, category, and the general BLARK matrix.

The concepts of size and quality

This BLARK introduces the notions of size (small/medium/large) and quality (low/medium/high) - and also license, which is discussed separately. This means that most, but not all, of the resources will be scored according to these three notions. However, their relative importance will vary across resources. Regarding size, we have considered it an important feature because having enormous amounts of data to be able to train large models is an essential aspect of LTs nowadays. The more text or speech data, the better, and even though this has always been the case, size is an even more determining factor today than two decades ago. We will provide some guidelines to help decide what is small, medium, and large depending on the type of data or resource. Following our methodology, these guide parameters were created by surveying existing resources in different languages.

The same applies to the concept of quality, whose components will also vary depending on the resource. By quality, we generally mean the process by which the resource was created, its performance on the task, or the amount of annotated information. For instance, whether the resource was taken directly from the internet without revision or cleaning, whether it has been manually or automatically annotated, or whether the annotations cover multiple linguistic aspects or just a few. The procedure to create the guidelines to measure quality was the same as for quantity, and in both cases, we ultimately leave it to the researcher's judgment, which should always take precedence regardless of the indications provided to establish the approximate measurements.

The concept of license

Legal considerations in the field of LTs are as important as they are complex. A detailed examination of these falls outside the scope of the present BLARK proposal, with the exception of issues related to the licensing of resources. The license category is common to all the resources, and weighted with 30% in all cases, as we believe that its value is of equal importance throughout. There are two main reasons why we have placed so much importance on the type of license, such that it is the only feature common to all the resources. Firstly, nowadays, it is essential that all the basic LT resources, especially those for minority languages, are free to be used in research, to advance the state of the art, and industry, to provide users with good quality products, particularly in contexts where big corporations have no market interests. Resources that are not released, are not publicly available, or that are priced too high, have much less value for the LT community in an under-resourced context. The second reason is that this may be the characteristic that varies the most between research and companies. When a resource is intended to be used for research purposes, it is usually easier to acquire than when it is intended to be used for commercial purposes. Hence, the type of license establishes a fundamental difference between the needs of academia and industry. The licensing feature is divided into three levels (closed, restricted, and open), each having

different weights (15%, 30%, and 100%, respectively). Given the wide variety of licenses and aspects that can be taken into consideration, below we provide some general indications and examples that would fall into each of the levels.

- **Closed:** by closed licenses, we refer to all those licenses that regulate public resources that can only be accessed via web queries (e.g. dictionaries like RAG[5] or corpora like the CORGA), commercial products that are open to the public on a limited basis, such as machine translation systems that can only be used free of charge with a limited number of characters (e.g. Google translate or Deepl. This situation is similar to tools like GPT chat, that can only be used after registration and whose training information is increasingly closed to the public. Finally, any closed commercial product would fall under this license.
- **Restricted:** open-source licenses that preserve the rights of developers to benefit from the commercial use of their work (e.g. the Commons Clause), or copy-left licenses that require all modified and extended versions of an open-source program to be free (e.g. GPL or AGPL licenses). These restrictions are especially problematic for small companies that want to commercialize their products.
- **Open:** open-source licenses that allow the distribution and modification of the resource, regardless of copyright. A typical license type that would fall within this category is Apache 2.0.

Given that the license section is common to all the resources reported in the BLARK, these considerations will not be repeated for each subcategory, as the same criteria apply to all.

The concept of resource

What is meant when we talk about a "resource" is the fourth issue that deserves clarification before moving into the specific subcategories. Of course, a given language may have, for instance, several reference corpora, or translation models, and each of those would represent an individual resource under the general understanding of the concept of resource. However, for this BLARK, we blur the concept of resource meaning all the available resources in a given subcategory, trading off between the capacity to collect all the information possible and the capacity to evaluate the degree of development of a language with the least information possible. We wanted the BLARK to be an easy-to-use tool that could be completed in a timely manner. Consequently, rather than having to introduce the information for each of the available resources within a particular category, we ask the BLARK user to provide an approximate judgment of their availability. For some resources, like corpora, it is possible to judge their total size when there are multiple resources for the same purpose (e.g. several data collections that can be used as a macro corpora). For other resources, like models or tools, we encourage the researcher to fill in the BLARK with the best performing one, or approximate the capabilities of all the available ones. In cases where it is not possible to find the exact characteristics of size, quality, or license, the table may be completed in an approximate manner. Ultimately, we insist that this BLARK was not designed as a resource catalog or a detailed report on the state of the art of a language, as there already exist very good tools and reports that fulfill these purposes within the ELE (such as the ELG or the ELE reports) or other sources.

In the following subheadings, we describe the subsections of the BLARK in detail. First, within the cross-cutting resources category, and then, within the LT tasks category.

[5] Dictionary of the Royal Galician Academy

4.1. Cross-cutting resources

Cross-cutting resources are those that can be used for many different purposes and in many different tasks (such as data pre-processing, model fine-tuning, etc.). It is divided in four subcategories, which we will explain in detail in what follows: corpora (70%), lexical resources (10%), NLP tools (10%), and language models (10%).

4.1.1. Corpora

The corpora subcategory covers those basic textual resources that can be used for several different tasks. It amounts to 70% of the cross-cutting resources category, and it is divided in annotated corpus, reference corpus, and macro corpus, as shown in Table 3 (which is a repeated version of Table 2). In turn, each of these resources is assessed with respect to their size, quality, and license.

Resource	**Size 30%**	**Quality 40%**	**License 30%**
Annotated corpora **20%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
	Size 40%	**Quality 30%**	**License 30%**
Reference corpus **30%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
Macro corpus **50%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	

Table 3: Cross-cutting resources: corpora. Note that corpora are worth 70% of the cross-cutting resources category.

The approximate guidelines to assess these resources are the following:

- **Annotated corpus:** refers to any kind of annotated collection of texts that can be used to train models for specific tasks (e.g. NERC, POS-tagging, etc.). To define these approximate guidelines we have revised annotated corpora of different low-resource and medium-resource languages such as Galician (CTG corpus), German (MTP), or Spanish (IULA Spanish LSP Treebank).
  - **Size:**
    - **Small:** <100k tokens
    - **Medium:** 100k-1M tokens
    - **Large:** >1M tokens
  - **Quality:** quality is defined with respect to the number and quality of linguistic annotations for relevant linguistic categories it contains.
    - **Low:** it covers one or two annotation categories, or they have low quality
    - **Medium:** it covers a variety of annotation categories and its quality is good enough to fine-tune small model-tasks.
    - **High:** it covers most or all the relevant linguistic categories and its quality is excellent in order to train good-performing models.

- **Reference corpus:** refers to a collection of texts that contains the different contemporary linguistic varieties of a language, it is representative, linguistically revised, and is designed to provide comprehensive information about language use in different domains and registers. To define these guidelines we have revised reference corpora of different languages such as Galician (CORGA), Spanish (CREA), or German (Schweizer Textkorpus).
  - **Size:**
    - **Small:** <10M tokens
    - **Medium:** 10M-100M tokens
    - **Large:** >100M tokens
  - **Quality:**
    - **Small:** it is not representative of the language, it does not cover most text types and domains.
    - **Medium:** it is quite representative of the language, it covers some text types and domains
    - **High:** it is fully representative of the language, it covers a wide range of text types, domains, and dialects.
- **Macro corpus:** refers to an enormous collection of texts. The difference with the reference corpus is that the macro corpus does not have to be representative of the linguistic varieties, nor is it as carefully curated as a reference corpus. It simply contains as much text as possible, even if it has not been checked by language professionals. To define these guideliness we have revised macro corpora of different languages like Galician (SLI GalWeb.1.0), Czech (SYN v4), Spanish (Compilation of Large Spanish Unannotated Corpus), Basque (EusCrawl), or the multilingual corpus OSCAR.
  - **Size:**
    - **Small:** <100M
    - **Medium:** 100M-1000M
    - **Large:** >1000M tokens
  - **Quality:**
    - **Low:** it has not been revised, nor cleaned
    - **Medium:** it has been partially cleaned but not revised by professional linguists
    - **High:** it has been carefully cleaned and (partially) revised by professional linguists.

4.1.2. Lexical resources

The lexical resources subcategory covers those collections of words that are fundamental for many NLP tasks. It amounts to 10% of the cross-cutting resources category, and it is divided into annotated lexicons and dictionaries, both having the same importance, as shown in Table 4. An annotated lexicon is assessed with respect to its size, quality, and license, while dictionaries are only assessed with respect to size and license, as prescriptive academic dictionaries are assumed to have a high-quality standard.

Resource	**Size 70%**		**License 30%**
Dictionaries **50%**	Small **20%** Medium **60%** Large **100%**		Closed **15%** Restricted **30%** Open **100%**
	Size 30%	**Quality 40%**	**License 30%**
Annotated lexicons **50%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

Table 4: Cross-cutting resources: lexical resources. Note that lexical resources are worth 10% of the cross-cutting resources category.

The approximate guidelines to assess these resources are the following:

- **Annotated lexicons:** refers to a list of the words (as exhaustive as possible) accompanied by a set of tags that add information about the word (lemma, grammatical category, morphological information, semantic information, polarity, etc.). They can be wide-covering or more specific in nature, such as polarity lexicons (e.g. the Sentiment lexicon for Irish or the Bootstrapped Lexicon of German Verbal Polarity Shifters), which are devised for the more specific task of sentiment analysis. When there are multiple annotated lexicons in a language, we encourage researchers to assess their size and quality globally. These ranges were established with respect to existing monlingual lexicons such as the esLEX or multilingual lexicons like the MCR.
  - **Size:**
    - **Small:** <1K entries
    - **Medium:** 1K-30K entries
    - **Large:** >30K entries
  - **Quality:** quality is defined with respect to the number, diversity, and quality of linguistic annotations it contains.
    - **Low:** the annotations cover a few categories or have low quality (as they may not be precise or contain errors).
    - **Medium:** the annotations include a relatively wide range of categories with acceptable quality.
    - **High:** the annotations cover the vast majority of relevant word information and the quality is excellent.
- **Dictionaries:** a dictionary is an academic or prescriptive source collecting the words of a language together with their definition and other relevant information (pronunciation, synonyms, etc.). The dictionaries for Galician (RAG), Basque (Basque General Dictionary) or Spanish (COES) have been used as a reference to create these size guidelines.
  - **Size:**
    - **Small:** <10K entries
    - **Medium:** 10K-100K entries
    - **Large:** >100K entries

4.1.3. NLP tools

Resource		**Existence 70%**	**License 30%**
POS tagging **20%**	Tokenization **5%**	Yes/No **100%**	Closed **15%** Research **30%** Open **100%**
Parsing **20%**	Lemmatization /word segmentation **5%**		
NERC **20%**	Word sense disambiguation **5%**		
Language identification **20%**	Coreference resolution **5%**		

Table 5: Cross-cutting resources: NLP tools. Note that NLP tools are worth 10% of the cross-cutting resources category.

The NLP tools subcategory covers basic tools that are necessary to pre-process and analyze data for a variety of tasks. It amounts to 10% of the cross-cutting resources category, and it contains a list of the most frequent and necessary NLP tools. Despite the fact that the range of possible NLP tools of interest is very vast, we have selected eight basic ones that are most useful and essential, shown in Table 5. The first four (i.e. POS tagging, parsing, NERC and language identification) have received more importance because they are key to pre-processing and cleaning data. Although tokenization is a fundamental pre-processing step for almost any LT task, most current tokenization methods are language-independent, which is why it has not received so much weight in the BLARK. The resources in this category are assessed with respect to two aspects: whether the tool exists with sufficient quality, and their license. We should emphasize that this subcategory is designed to cover tools that function using deep learning methods.

4.1.4. Language Models

The last subcategory within cross-cutting resources is language models, and it amounts to 10% of cross-cutting resources. Large pre-trained language models (also known as foundation models) have caused a complete paradigm shift in LTs, achieving state-of-the-art results in many NLP and LT tasks (Min et al., 2021b). This category is divided into word embeddings, monolingual autoencoder models, monolingual autoregressive models, multilingual autoencoder models, and multilingual autoregressive models, each associated with a different weight, as shown in Table 6. We have given more value to autoencoder than autoregressive models because autoregressive models demand, even for fine-tuning, much larger amounts of data than autoencoder models. Thus, its adaptation to minoritized languages is more challenging than autoencoder models. There are, of course, other types of models, such as encoder-decoder or sequence-to-sequence (e.g. T5 (Raffel et al., 2020)). These have not been included in this subsection because they play a more important role for specific LT tasks.

Resource	**Size 70%**		**License 30%**
Embeddings **10%**	**Vocabulary size 50%**	**Training corpus 50%**	
	Small **20%** Medium **60%** Large **100%**	Small **20%** Medium **60%** Large **100%**	Closed **15%** Research **30%** Open **100%**
	Nº Parameters 40%	**Training corpus 60%**	
Autoencoder (monolingual) **30%**	Small **20%** Medium **60%** Large **100%**	Small **20%** Medium **60%** Large **100%**	Closed **15%** Research **30%** Open **100%**
Autoregressive ´ (monolingual) **30%**	Small **20%** Medium **60%** Large **100%**	Small **20%** Medium **60%** Large **100%**	Closed **15%** Research **30%** Open **100%**
	Nº Parameters 40%	**% of the language in training data 60%**	
Autoencoder (multilingual) **15%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Research **30%** Open **100%**
Autoregressive (multilingual) **15%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Research **30%** Open **100%**

Table 6: Cross-cutting resources: language models. Note that language models are worth 10% of the cross-cutting resources category.

Autoencoder models are trained using the encoder part of the original Transformer architecture (Vaswani et al., 2017). They are able to predict the next word by collecting bidireccional information of the sequence. The BERT (Devlin et al., 2019) family is the most representative example of autoencoder models. These models can be fine-tuned to carry out different classification tasks such as NERC or POS-tagging, achieving state–of-the-art results. Autoregressive models are trained using the decoder part of the Transformer architecture to predict the next word given all the previous words in the sequence. The most famous examples of autoregressive models are the different versions of GPT (Radford et al., 2018), GPT-2 (Radford et al., 2019), and GPT-3 (Brown et al., 2020). This type of models can be fine-tuned for different tasks, particularly those that require language generation, although not exclusively (Min et al., 2021b). Providing evaluation criteria for language models is very challenging for two main reasons. First, at the speed at which new models are emerging these days, the guidelines provided below with respect to current state-of-the-art models will surely be outdated soon. Secondly, the data and architecture used are not always public, making it difficult to provide relevant values.

The approximate guidelines to assess these resources have been defined based on Min et al. (2021a) paper:

- **Word embeddings:** refers to “dense, distributed, fixed-length word vectors, built using word co-occurrence statistics as per the distributional hypothesis” (Alt et al., 2019, p.2). To provide the following approximate guidelines, we have considered the word embeddings of different languages like Catalan, Galician, Spanish, Portuguese or German[6].
  - **Size:**
    - **Vocabulary size:**
      - **Small:** < 500k tokens

[6] Source: http://vectors.nlpl.eu/repository/

        - **Medium:** 500k - 1M tokens
        - **Large:** > 1M tokens
    - **Training corpus size:**
        - **Small:** < 3B tokens
        - **Medium:** 3B tokens < 10B tokens
        - **Large:** > 10B tokens

- **Monolingual autoencoder model:** to provide the following approximate guidelines, we have consulted the characteristics of BERT models of different languages like English, Galician, Basque (Agerri et al., 2020b) or Portuguese (Souza et al., 2020).
  - **Size:**
    - **Nº parameters:**
        - **Small:** <100M parameters
        - **Medium:** 110M parameters
        - **Large:** > 110M parameters
    - **Training corpus size:**
        - **Small:** <200M tokens
        - **Medium:** 200M - 500M tokens
        - **Large:** > 500M tokens

- **Monolingual autoregressive model:** to provide the following approximate guidelines, we have reviewed the GPT, GPT2 small and base and GPT3 English models (Min et al., 2021a). We did not consider GPT models in other languages as there are few languages trained with this type of architecture and because the training parameters and corpus are not easily available.
  - **Size:**
    - **Nº parameters:**
        - **Small:** <124M parameters
        - **Medium:** 124M-300M parameters
        - **Large:** >300M parameters
    - **Training corpus size:**
        - **Small:** <800M tokens
        - **Medium:** ca. 12B tokens
        - **Large:** > 12B tokens

- **Multilingual autoencoder model:** to provide the following approximate guidelines we have used mBERT-distilled and mBERT-base (Devlin et al., 2018) as reference.
  - **Size:**
    - **Nº parameters:**
        - **Small:** < 110M parameters
        - **Medium:** 110M - 177M parameters;
        - **Large:** > 177M parameters
    - **% of the language in the training data:**

        - **Small:** < 0.17%
        - **Medium:** < 2.8%
        - **Large:** > 2.8%

- **Multilingual autoregressive model:** to provide the following approximate guidelines we have used multilingual GPT (Shliazhko et al., 2022) as reference.
  - **Size:**
    - **Nº parameters:**
      - **Small:** < 1.3B parameters
      - **Medium:** 1.3B < 1.5B parameters
      - **Large:** > 1.5B parameters
    - **% of the language in the training data:**
      - **Small:** < 0.01%
      - **Medium:** 0.01% - 0.60%
      - **Large:** > 0.60%

4.2. LT tasks

LT tasks represent the most important category of the BLARK, amounting to a 60% of the final score. It is divided into five main subcategories. First, the three basic tasks that, in our view, any minoritized language should have as a priority: speech synthesis (4.2.1), speech recognition (4.2.2) and machine translation (4.2.3). Then, other LT tasks (4.2.4), that present further challenges and are often hardly supported in minoritized languages, at least using deep learning methods. These include: grammatical error correction, summarization, sentiment analysis, fact-checking, and dialog systems. Finally, the last subcategory of the BLARK is benchmarking (4.2.5), which condenses, in a very general way, the importance of having sufficient evaluation materials to analyze the quality of the different LT systems. Most of these subcategories, although not all, will be evaluated according to what are, in our view, the two main axes of LT development: corpora (i.e. the availability of data) and models (the existence of trained architectures that can perform the task). As already noted, to complete the information about the size and quality features of the corpora for each task, we encourage researchers to take into account, whenever possible, all the available corpora for a particular task instead of just an individual resource, approximating a total size. With respect to model quality, we encourage researchers to describe the features of the best-performing model. We provide some guidelines with respect to the most representative metrics for each task, although this should be taken with caution, as automatic metrics are simply approximate indicators of the quality of a model, but are not always reliable.

4.2.1. Speech synthesis

Speech synthesis (also known as text-to-speech (TTS) can be defined as the task of synthesizing intelligible and natural speech from text (Tan et al., 2021). Although the resources that could be used to train speech models (or any other model) are very heterogeneous, for reasons of clarity, this and the following LT tasks will be evaluated in terms of corpus and models, as shown in Table 7.

Resource	**Size 40%**	**Quality 30%**	**License 30%**
Corpus **70%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
Model **30%**		**Quality 70%**	**License 30%**
		Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

Table 7: LT tasks: speech synthesis. Note that speech synthesis is worth 20% of the LT tasks category.

The approximate guidelines to assess these resources are the following:

- **Corpus**
  - **Size:**
    - **Small:** < 4 hours
    - **Medium:** 4 - 10 hours
    - **Large:** > 10 hours
  - **Quality:** the key features to evaluate the quality of the corpus are the recording quality and the text balancing (i.e. phonetic distribution, representativeness, etc.).
    - **Low:** recorded by non-professionals and the text is phonetically unbalanced.
    - **Medium:** the corpus has been recorded by semi-professionals and the text is quite balanced.
    - **High:** recorded by professionals and the text is well-balanced.
- **Model:**
  - **Quality:** model quality is often evaluated using the MOS metric (Mean Opinion Score).
    - **Low:** < 3 MOS score
    - **Medium:** 3 < 4 MOS score
    - **High:** > 4 MOS score

4.2.2. Speech recognition

Automatic speech recognition (ASR) is the task that facilitates the recognition and translation of spoken language into text by a machine (Rista and Kadriu, 2020). As shown in Table 8, this task is evaluated in terms of corpus and models.

Resource	**Size 40%**		**Quality 30%**	**License 30%**
	Nº hours 60%	**Nº speakers 40%**	Low **20%**	Closed **15%**
Corpus **70%**	Small **20%** Medium **60%** Large **100%**	Small **20%** Medium **60%** Large **100%**	Medium **60%** High **100%**	Restricted **30%** Open **100%**
Model **30%**			**Quality 70%**	**License 30%**
			Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

Table 8: LT tasks: speech recognition. Note that speech recognition is worth 20% of the LT tasks category.

The approximate guidelines to assess these resources are the following:

- **Corpus:**
  - **Size:**
    - **Nº of hours:**
      - **Small:** < 500 hours
      - **Medium:** 500 - 2k hours
      - **Large:** > 2k hours
    - **Nº of speakers:**
      - **Small:** < 200 speakers
      - **Medium:** 200 - 5k speakers
      - **Large:** >5k speakers
  - **Quality:**
    - **Low:** automatically aligned corpus
    - **Medium:** automatically aligned corpus with a quality filter
    - **High:** manually transcribed corpus
- **Model:**
  - **Quality:** model quality is often evaluated using the WER metric (Word Error Rate)
    - **Low:** > 20% WER score
    - **Medium:** 20% - 10% WER score
    - **High:** < 10% WER score

4.2.3. Machine translation

Machine translation (MT) can be defined as the use of computerized systems to transform a text written in a source language into a text written in a target language, generating a translation (Forcada, 2020). In this BLARK, a key element for the evaluation of the development of MT systems is the number of languages that can be translated to and from (i.e. translation pairs). Given that many minoritized languages have hardly developed a few language pairs with sufficient quality using neural machine translation systems (NMT), only a maximum of 4 language pairs are evaluated per language. Each language pair will have the same weight. As shown in Table 9, for each translation pair, this task is evaluated in terms of the available corpora and models.

Language pair	Resource	Size 40%	Quality 30%	License 30%
For a maximum of 4 pairs (=25% each)	Corpus **70%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
	Model **30%**		**Quality 70%**	**License 30%**
			Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

Table 9: LT tasks: machine translation. Note that machine translation is worth 20% of the LT tasks category.

The approximate guidelines to assess these resources are the following:

- **Corpus:**
  - **Size:**
    - **Small:** < 2M parallel sentences
    - **Medium:** 2M - 10M parallel sentences
    - **Large:** > 10M parallel sentences
  - **Quality:** the most common criteria to evaluate corpus quality will be font origin and sentence alignment. The font is important because texts crawled from the internet, especially from unknown pages, are linguistically poor or even machine translated, and have many basic mistakes. Sentence alignment is a crucial preprocessing step. Usually, automatic alignments without any kind of revision generate parallel corpora with empty lines and misalignments whose detection and correction are not trivial.
    - **Low:** unknown font, crawled from unknown or not curated internet pages, automatically aligned with no revision/cleaning.
    - **Medium:** human translations, automatically aligned with revision/cleaning.
    - **High:** aligned human (professional) translations (e.g. translation memories)
- **Model:**
  - **Quality:** model quality is often evaluated using the BLEU metric (Bilingual Evaluation Understudy) (Papineni et al., 2002).
    - **Low:** < 30 BLEU score
    - **Medium:** 30 - 70 BLEU score
    - **High:** > 70 BLEU score

Machine translation has been a difficult task to evaluate, because, being the most socially widespread LT, there exist many online commercial translation services (e.g. Google Translate, DeepL, Yandex, etc.) that include a varying degree of minority languages. However, as these services are private, it is difficult to evaluate them, and are not directly contributing with resources that can be used by the LT community. For these reasons, they have not been included in this BLARK, although their positive impact on digital equality for minoritized languages is undeniable. In connection with the foregoing, it is also true that there exist multilingual models that integrate many minoritized languages (e.g. M2M (Fan et al., 2020) and NLLB (Costa-jussà et al., 2022)). We encourage users to consider these types of models when they have been fine-tuned for a particular language pair, resulting in a system with sufficient quality.

4.2.4. Other LT tasks

In this subsection we have integrated several tasks that are a central part of LTs but that are also scarcely covered even by high-resource languages other than English (e.g. German or Spanish). Their degree of development in a minoritized language will be positively valued, but their absence will not have such a negative impact on the final BLRAK scoring. These tasks are: grammatical error correction (4.2.4.1), summarization (4.2.4.2), sentiment analysis (4.2.4.3), fact-checking (4.2.4.4) and dialog systems (4.2.4.5). As in previous cases, the assessed resources are corpora and model performance, although it has been difficult to establish quantitative ranges with respect to corpus sizes due to the variability within existing resources. We have decided to establish a general guideline in this respect, that applies to all the corpora for the LT tasks described below. To avoid repetition, we will not include these size guidelines for each of the corpora covered in this subsection. General corpora size guidelines:

- **Small:** the available data is not enough data to fine-tune an existing model with sufficient quality
- **Medium:** there is enough data to fine-tune an existing model obtaining relatively good results
- **Large:** there is enough data to train a model from scratch.

4.2.4.1. Grammatical error correction

Grammatical error correction (GEC) is the task of automatically detecting and correcting errors in text. These errors can be orthographic, grammatical, semantic, or related to style or tone (Bryant et al., 2022). The biggest difficulty to train good grammatical error correction models is the absence of big-enough high-quality data, even for English. As shown in Table 10, this task is evaluated in terms of corpus and models.

Resource	**Size 40%**	**Quality 30%**	**License 30%**
Corpus **70%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
Model **30%**		**Quality 70%**	**License 30%**
		Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

Table 10: LT tasks/Other LT tasks: grammatical error correction. Note that grammatical error correction is worth 20% of the Other LT tasks subcategory.

The approximate guidelines to assess these resources are the following (note that corpus size guideless are described in Section 4.2.4):

- **Corpus:**
  - **Quality:** the quality of the corpus is often evaluated depending on annotation consistency and the diversity of errors that it contains.
    - **Low:** the corpus does not have a wide variety of annotated errors. For example, it only has orthographic errors but does not consider other types (e.g. semantic, grammatical, etc.).

    - **Medium:** the corpus covers all basic errors (spelling, grammar, etc.)
    - **High:** the corpus has a huge variety of annotated errors. It does not only covers basic mistakes (spelling, grammar), but also semantic, and style and tone.

- **Model:**
  - **Quality:** model quality is often evaluated using the GLEU metric (Generalized Language Understanding Evaluation) (Napoles et al., 2016)
    - **Low:** < 0.5 GLEU score
    - **Medium:** 0.5 - 0.8 GLEU score
    - **High:** > 0.8 GLEU score

4.2.4.2. Summarization

Text summarization is the method "to reduce the source text into a compact variant, preserving its knowledge and actual meaning" (Mridha et al., 2021, 1). Automatic text summarization is an LT task that can be divided into two main classes depending on how the task is tackled: extractive and abstractive. Extractive text summarization is the most used and the most reliable strategy these days. It takes sentences from the original input text to make the summary, so, as the model does not generate new data to make the summary, correctness is assured linguistically-wise and content-wise. Abstractive summarization class generates new data to make the summary, being riskier, but also, the summary generated tends to be more fluent than the one generated using extractive summarization. To keep the basic nature of this BLARK, we will not distinguish between these two main strategies by giving them different weights, and instead, encourage users to complete the BLARK with whichever system they deem appropriate. As shown in Table 11, this task is evaluated with respect to corpora and models. Nonetheless, some differences apply with respect to the corpus. The two summarization strategies need parallel corpora containing the original text and its summarization. Yet, extractive training data needs to be labeled in the summarization part to be able to recognize the important information in each summary, whereas the abstractive class only needs large amounts of parallel data without any type of labeling. Model quality is evaluated following the same criteria regardless of the type of strategy.

Resource	**Size 40%**	**Quality 30%**	**License 30%**
Corpus **70%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
Model **30%**		**Quality 70%**	**License 30%**
		Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

Table 11: Summarization table evaluated according to corpora and models. LT tasks/Other LT tasks: summarization. Note that summarization is worth 20% of the Other LT tasks subcategory.

The approximate guidelines to assess these resources are the following (note that corpus size guideless are described in Section 4.2.4):

- **Corpus:**

  - **Extractive corpus quality:** corpus quality is often evaluated by how it was labeled.
    - **Low:** sentences automatically labeled without revision, covering a small variety of features.
    - **Medium:** sentences automatically labeled but revised, covering a medium variety of features.
    - **High:** sentences manually labeled, covering a large variety of features.
  - **Abstractive corpus quality:** corpus quality is often evaluated by how the summary was created.
    - **Low:** automatically generated without human revision.
    - **Medium:** automatically generated but manually revised.
    - **High:** manually generated summaries.
- **Model:**
  - **Model quality:** model quality is often evaluated using the ROUGE F1 score (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004).
    - **Small:** < 0.2 ROUGE F1 score
    - **Medium:** 0.2 < 0.4 ROUGE F1 score
    - **High:** > 0.4 ROUGE F1 score

4.2.4.3. Sentiment Analysis

Sentiment analysis is the task of determining the polarity of a given text, identifying and tagging data according to a positive, negative or neutral sentiment (Umar et al., 2018). There are several possible sentiment analysis tasks (e.g. document level classification, sentence level classification, aspect-based sentiment analysis, etc.), although this BLARK only covers a general view of sentiment analysis. As shown in Table 12, this task is evaluated with respect to corpora and models.

Resource	**Size 40%**	**Quality 30%**	**License 30%**
Corpus **70%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
Model **30%**		**Quality 70%**	**License 30%**
		Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

Table 12: LT tasks/Other LT tasks: sentiment analysis. Note that sentiment analysis is worth 20% of the Other LT tasks subcategory.

The approximate guidelines to assess these resources are the following (note that corpus size guideless are described in Section 4.2.4):

- **Corpus:**
  - **Quality:** corpus quality is often evaluated by the availability of tags (amount and quality) in different domains.

    - **Low:** a small variety of unique entities tagged and a small number of tokens significantly different in a few domains.
    - **Medium:** a medium variety of unique tagged entities in different domains
    - **High:** a large amount of unique entities tagged in many different domains

- **Model:**
  - **Quality:** although the best metric to measure model quality can vary depending on the task (e.g. precision, recall, F1, or accuracy), we take accuracy as a good indicator of model performance.
    - **Small:** < 50% accuracy
    - **Medium:** 50% < 80
    - **High:** > 80% accuracy

4.2.4.4. Fact checking

Fact-checking is the task of "assessing whether claims made in written or spoken language are true" (Guo et al., 2022, 179). One of the main difficulties to train good fact-checking models is the absence of rich annotated corpora in different domains (Hanselowski et al., 2019). This is already quite challenging for high-resource languages and is aggravated in low-resource scenarios. As shown in Table 13, this task is evaluated in terms of corpus and models.

Resource	**Size 40%**	**Quality 30%**	**License 30%**
Corpus **70%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
Model **30%**		**Quality 70%**	**License 30%**
		Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

Table 13: LT tasks/Other LT tasks: fact checking. Note that fact checking is worth 20% of the Other LT tasks subcategory.

The approximate guidelines to assess these resources are the following (note that corpus size guideless are described in Section 4.2.4):

- **Corpus:**
  - **Quality:** corpus quality can be evaluated by the variety of labels, fake news and domains
    - **Low:** there is not a wide variety of labeled data in basic domains.
    - **Medium:** there is a wide variety of labeled data in several domains.
    - **High:** there is a large variety of labeled data claims in many domains.
- **Model:**
  - **Quality:** although there are several possible metrics to evaluate model performance (e.g. precision, recall or F1), we provide some guidelines using F1:
    - **Low:** < 50% F1

* **Medium:** 50% - 80% F1
* **High:** > 80% F1

4.2.4.5. Dialog systems

Dialog systems integrate a variety of subtasks (e.g. natural language understanding, classification, dialog management, etc.) depending on the architecture of the system and the system's goal, making it difficult to evaluate its development using the same criteria as for the rest of the tasks while keeping the BLARK as simple as possible. After much analysis, it was divided into two broad categories: usability and corpora, as shown in Table 14.

<table>
<tr><th>Type</th><th colspan="2">Can you interact
in your own language with...</th><th>Answers</th></tr>
<tr><td>Usability 30%</td><td colspan="2">Mobile assistants 14,3%
LLM-based general chatbots and Q&A systems 14,3%
Smart speakers 14,3%
Public administration dialog systems 14,3%
Frequent e-commerce dialog systems 14,3%
Health applications dialog systems 14,3%
Other task-oriented dialog systems 14,3%</td><td>Barely 20%
Partially 60%
Mostly 100%</td></tr>
<tr><td>Corpora 70%</td><td>Corpora size 20%</td><td>Corpora Quality 50%</td><td>License 30%</td></tr>
<tr><td rowspan="2">Corpora for domain specific dialog systems 40%</td><td rowspan="2">Small: 20%
Medium: 60%
Large: 100%</td><td>Annotations 50%</td><td>Number of domains 50%</td><td rowspan="2">Closed 15%
Restricted 30%
Open 100%</td></tr>
<tr><td>Low 20%
Medium 60%
High 100%</td><td>Few domains 20%
Some domains 60%
Many domains 100%</td></tr>
<tr><td></td><td>Corpora size 40%</td><td>Corpora Quality 30%</td><td>License 30%</td></tr>
<tr><td>Corpora for general generative dialog systems 60%</td><td>Small 20%
Medium 60%
Large 100%</td><td>Low 20%
Medium 60%
High 100%</td><td>Closed 15%
Restricted 30%
Open 100%</td></tr>
</table>

Table 14: LT tasks/Other LT tasks: dialog systems. Note that systems is worth 20% of the Other LT tasks subcategory.

Usability refers to the availability of dialog system tools (e.g. conversational assistants, chatbots, etc.) for a particular language. These tools are now completely integrated into people's daily lives, but speakers of minoritized languages cannot often use them in their own language. As shown in Table 14, the usability category contains a list of these services, common whose implementation for minoritized languages is evaluated using three levels:

- **Usability:**
  - **Barely:** this tool exists in my language, but in very specific situations and cannot be normally used.
  - **Partially:** this tool exists in my language and can be normally used, but there is not a variety of different options.

  - **Mostly:** this tool exists in my language and there are a wide variety of different options.

Dialog systems can be divided into two main categories: domain-specific, which entails models trained for a very specific task or problem (e.g. movie ticket booking), and open domain, which entail models that can interact with the user without any domain restrictions (Ni et al., 2022). Depending on the category, the required training corpus varies in size and requirements. For example, in domain-specific tasks, corpus quality is more important than corpus size, and vice versa. This is because domain-specific tasks are limited to a narrower set of interactions, making it more important that the corpus has a wide variety of labels to capture detailed interaction types. Contrarily, open-domain dialog systems need enormous amounts of corpora to incorporate many different domains and interaction types.

The approximate guidelines to assess these resources are the following (note that corpus size guideless are described in Section 4.2.4):

- **Domain-specific corpus:**
  - **Quality annotations:** corpus quality is often evaluated with respect to its annotations (e.g. entities, slots, intents or dialog acts) and number of domains.
    - **Low:** covers basic features (e.g. entities or intents) but not others (e.g. slots or dialog acts).
    - **Medium:** covers all the important information necessary to train medium dialog systems, including different intents, slots, entities and dialog acts.
    - **High:** covers all the important information necessary to train good domain specific dialog systems, it can be used to train basic task-oriented systems and also more specific and assorted ones.
  - **Quality nº of domains:** nº of domains and subdomains for which there are available corpora:
    - **Few domains:** <10
    - **Some domains:** 10-20
    - **Many domains:** > 20
- **Open-domain corpus:**
  - **Quality:** corpus quality will be evaluated with respecto to the font and the cleaning process.
    - **Low:** unknown font, not revised nor cleaned.
    - **Medium:** labeled font automatically revised and cleaned.
    - **High:** corpus manually revised and cleaned.

4.2.5. Benchmarking

This last subsection covers, in a very general way, the importance of having sufficient evaluation materials to analyze the quality of the different LT systems developed for each of the LT tasks. Even though this is not an LT task per se, but rather, a subtask, being able to evaluate the quality of a system, and compare it with other similar systems, is essential. There are many low-resource languages that do not have gold-standard datasets to evaluate specific tasks or are not included in multilingual benchmarks that allow researchers to compare the results of the systems with other languages. Some examples of these type of resources could be FLORES (Goyal et al., 2021) or TATOEBA (Tiedemann, 2020) for machine translation or the

natural language understanding benchmark SuperGLUE (Wang et al., 2019), among many others. To evaluate this last section, the user will have to answer two different questions (shown in table 15). The first question covers widely used evaluating resources that facilitate comparisons between languages. The second question refers to evaluating resources that were developed individually for the language, and that allow to evaluate the quality of LT systems but not to directly compare them with other languages. As with the previous categories, not answering implies a total lack of evaluation resources.

Benchmarking 10%		
Taking into account the LT tasks mentioned above...		
Resource	**Question**	**Answer**
Widely used evaluation resources **60%**	Is your language present in mainstream benchmarks or evaluation resources/datasets?	Barely **20%** Partially **60%** Mostly **100%**
Evaluation materials of its own **40%**	Does your language have sufficient evaluating resources of its own?	

Table 15: LT tasks: benchmarking. Note that benchmarking is worth 10% of the LT tasks category.

1. **Is your language present in mainstream benchmarks or widely used and standardized evaluation resources/datasets?**
   - **Barely:** the language is present for one or two specific tasks.
   - **Partially:** the language is present for some tasks but not for others.
   - **Mostly:** the language is present for all the previously mentioned LT tasks in this BLARK.
2. **Does your language have sufficient evaluating resources of its own?**
   - **Barely:** the language has evaluating resources for some basic LT tasks (e.g. machine translation, speech synthesis, or speech recognition) but not for most.
   - **Partially:** the language has evaluating resources for the basic LT tasks in different domains and also in some non-basic tasks (e.g. sentiment analysis or grammatical error correction).
   - **Mostly:** the language has sufficient evaluating resources for all the tasks mentioned in this BLARK.

5. GL-BLARK

In order to test the suitability of this BLARK proposal, we have piloted it using our knowledge about the situation of Galician. The detailed results for the Galician BLARK matrices (i.e. GL-BLARK) can be found in Appendix B.

In this BLARK, Galician has obtained a final score of 54,03% out of 100%. It has achieved a 62,29% in cross-cutting resources (i.e. 24,91% out of the 40% it contributes to the final BLARK), and a 48,54% in LT tasks (i.e. 29,12% out of the 60% it contributes to the final BLARK). Thus, according to this BLARK for minoritized languages, Galician is in a medium-low state of development. The first thing to notice from the scores obtained in these two general categories is that the degree of development of cross-cutting resources is better than that of LT tasks. This draws the picture of an underdeveloped language that is in much need of public support and private interest to provide its speakers with sufficient LTs of reasonable quality. However, in part thanks to the efforts put by the Nós Project, as well as other research

groups in the area of LTs, it is at a starting point from which there are interesting challenges and possibilities to work towards digital language equality. Let us take a closer look at the two broad categories of the BLARK and its subcategories.

With respect to cross-cutting resources, the degree of development of NLP tools is remarkable. One of the reasons for this positive development is that Galician is included in multilingual tools such as Freeling or Linguakit. In spite of the fact that the subcategories of lexical resources and corpora achieve a medium level of development, it should be noted that some of the most insteresting and useful resources are not fully open or can only be consulted via web queries (e.g. CORGA), or are of low quality (e.g. SLI GalWeb). Therefore, it would be a very significant step forward if all these resources were made openly available for their use by research initiatives and companies. In general, it would be necessary to invest in the creation of high-quality and varied corpora. Finally, regarding language models, Galician has several monolingual autoencoder models such as Bertinho (Calvo et al., 2021) and three BERT models (small, base and large), which is very significant plus point for a low resource language. Yet, the presence of Galician in multilingual models, both autoencoder (mBERT) and autoregressive (GPT) is very low. In general, more efforts are needed for the development of state-of-the-art language models, something which goes hand in hand with the availability of public corpora. These are areas where the Nós project is paying special attention, and for which new resources and models will soon be published in the corresponding repositories (see the Nós project GitHub, Zenodo, and HuggingFace pages).

With respect to LT tasks, the paradigm shift in speech technologies over the last year thanks to the Nós project is specially remarkable. In this short period of time, intensive work has been carried out on the collection of speech data. One of the major problematic issues noted by Sánchez and Mateo (2022), was that in Galician there was a greater development of text resources than speech ones. Nowadays, speech synthesis is, without any doubt, the most developed technology for Galician, while speech recognition seems to follow a parallel trend. The same occurs with respect to machine translation, a task in which until a few weeks ago when the Nós project launched the English-Galician and Spanish-Galician models, Galician was only present in multilingual models such as M2M or NLLB. However, in spite of the foregoing, there is a clear need for quality parallel corpora in these and other language pairs. On the other hand, the development of Galician in Other LT tasks (i.e. GEC, summarization, sentiment analysis, fact checking, and dialog systems) is almost non-existent. In some of them there is a small amount of corpus that would not even be sufficient to train a model (e.g. GEC and dialogue systems), while, to the best of our knowledge, there are no specific models to perform any of them. The case of dialogue systems is noteworthy. At a time when these tools are present in everyday life, there are practically no systems with which people can interact in Galician. With the exception of a limited number of public administration apps, there are no applications for individual use such as mobile applications or smart speakers. Finally, it is also important to highlight the poor results obtained in benchmarking. Galician is not included in standardized evaluation resources except some datasets as FLORES (Goyal et al., 2021) or TATOEBA (Tiedemann, 2020), nor has sufficient evaluation datasets of its own. Benchmarking is a fundamental tool to be able to evaluate the tasks' performance, hence it is crucial that there is more research in this area.

To the best of our knowledge, the scores obtained in the BLARK reflect the current degree of development of Galician quite well, also proving the BLARK's ability to capture LT development for a paradigmatic example of a minoritized language in Europe. Putting our BLARK to the test with Galician highlighted some of the problems and difficulties that, as we had foreseen, BLARK users can encounter. The most salient of these is the problem of approximating the values when there are multiple resources in one category, particularly when these resources have opposite characteristics. For instance, there are several available corpora for speech recognition, and they differ radically in the number of hours recorded, their quality and their license, such that the bigger corpus is not open yet (although it will be soon),

while the smaller one is already open. The strategy in this case has been to categorize the resources ”medium”, although it is not ideal, since it does not reflect the real situation with total fidelity. In any case, as already mentioned, this BLARK was not designed as a resource catalog or a detailed report on the state of the art of a language, as there already exist very good tools and reports that fulfill these purposes.

6. Conclusions and Limitations

Throughout this project, we have carried out an intense research work that has allowed us to analyze and categorize the basic resources, tools and tasks in different LT areas. We are aware of the fact that this is an initial proposal that will need constant revision and updating over time, especially after the speed at which LTs have been evolving in the past couple of years. Nonetheless, we hope to have provided minority languages with a tool to help determine their starting point, identify which key areas would need to be developed in order to cover the basic LTs, define areas of potential growth, and help funding bodies spot investment needs.

Despite the fact that we have considered different minority languages in our analyses, we are aware of the fact that the design of this BLARK is influenced by our knowledge and preconceptions about Galician and its gaps. Therefore, this tool needs to be tested in different contexts and languages in order to be validated. To achieve this, we hope to release a digital version of the BLARK with the support of ELE, such that different actors from research and industry can use it and provide us with feedback on possible improvements, errors, deficiencies, or inconsistencies. At the moment, we have been working on the development of a web-based BLARK with the intention of making this resource available and easy to use. We hope to release it in June 2023, upon having received the relevant feedback. On this website, any registered user will be able to fill in a BLARK for any language in the form of a user-friendly questionnaire that would list all the sections. Once the form has been filled in, the BLARK table can be downloaded, including the indicators of the degree of development for each category and subcategory. This way, the BLARK also becomes a tool of potential interest to survey the degree of development of European minoritized languages, and serve as a companion to the ELG Catalogue and the ELE language reports, as well as projects like the Digital Language Diversity Project DLDP. This project aims to preserve European linguistic diversity and has launched tools such as the Digital Language Equality Survival Kit, which allows speakers to assess the state of vitality of their language and learn about the kind of concrete actions and initiatives that can improve this level of vitality (Berger et al., 2018).

We have tried to bring the perspectives of academia and industry into the BLARK, although the ability to do this was limited by the higher goal of developing a BLARK matrix that can be completed in a timely manner. This entailed simplifying some aspects that we know are far from simple, but that conversely resulted in a BLARK that is general and flexible enough to adapt to new realities. One of the novelties that we find most appealing is the importance given to resource licensing, which is a key factor to advance research, but mostly, to provide industry with basic resources on top of which they can create end-user products. One aspect which could not be integrated into the BLARK, where industry and research differ, has to do with quality demands. Quality criteria for industry transferability are higher than those observed in research. On many occasions, resources that apparently have an optimum quality are not valid to be used in end-user products. Transferability is an important dimension to assess how LT advancements can be translated into real tools for minority language speakers.

Finally, we should insist that this BLARK was specifically designed to assess the degree of development of under-resourced languages. This is both an asset and a liability of the

proposal. It is an asset because minoritized languages are possibly the ones that need these type of evaluation tools the most, and that would benefit from being analyzed using criteria that are specifically geared. However, it is also a liability, as it is not the best tool to evaluate medium or high-resource languages, or compare different languages at different stages of development. In the future, it would be interesting to launch a large-scale BLARK that can be used for any type of language, regardless of its development in this area.

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. Give your text representation models some love: the case for Basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France, May 2020a. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.588.

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. Give your Text Representation Models some Love: the Case for Basque. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*, 2020b.

Christoph Alt, Marc Hübner, and Leonhard Hennig. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. *CoRR*, abs/1906.08646, 2019. URL http://arxiv.org/abs/1906.08646.

Klara Ceberio Berger, Antton Gurrutxaga Hernaiz, Paola Baroni, Davyth Hicks, Eleonore Kruse, Valeria Quochi, Irene Russo, Tuomo Salonen, Anneli Sarhimaa, and Claudia Soria. Digital language survival kit. the dldp recommendations to improve digital vitality. In *The Digital Language Diversity Project*. The Digital Language Diversity Project, 2018.

D. Binnenpoorte, F. De Vriend, J. Sturm, W. Daelemans, H. Strik, and C. Cucchiarini. A field survey for establishing priorities in the development of HLT resources for Dutch. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, May 2002a. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2002/pdf/252.pdf.

Diana Binnenpoorte, Catia Cucchiarini, Janienke Sturm, and Folkert Vriend. Towards a roadmap for human language technologies: Dutch-flemish. July 2002b.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. Grammatical error correction: A survey of the state of the art. *arXiv preprint arXiv:2211.05166*, 2022.

David Vilares Calvo, Marcos García González, and Carlos Gómez Rodríguez. Bertinho. *Procesamiento del lenguaje natural*, 66:13–26, 2021.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Iria de Dios-Flores, Carmen Magariños, Adina Ioana Vladu, John E. Ortega, José Ramom Pichel, Marcos García, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín-Diz, Manuel González González, Senén Barro, and Xosé Luis Regueira. The Nós Project: Opening routes for the Galician language in the field

of language technologies. In *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 52–61, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.tdle-1.6.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation, 2020.

Mikel L. Forcada. Building machine translation systems for minor languages: Challenges and effects. *Revista de Llengua i Dret // Journal of Language and Law*, (73), 2020.

Marcos Garcia. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3625–3640. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.281. URL https://aclanthology.org/2021.acl-long.281.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*, 2021.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. A richly annotated corpus for different tasks in automated fact-checking. *arXiv preprint arXiv:1911.01214*, 2019.

Steven Krauwer. Elsnet and elra: A common past and a common future. 1998. URL https://www.speech.kth.se/prod/blark/elsnet&elra.pdf.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Bente Maegaard, Steven Krauwer, Khalid Choukri, and Lise Damsgaard Jørgensen. The BLARK concept and BLARK for Arabic. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/521_pdf.pdf.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *CoRR*, abs/2111.01243, 2021a. URL https://arxiv.org/abs/2111.01243.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey, 2021b.

Muhammad F Mridha, Aklima Akter Lima, Kamruddin Nur, Sujoy Chandra Das, Mahmud Hasan, and Muhammad Mohsin Kabir. A survey of automatic text summarization: Progress, process and challenges. *IEEE Access*, 9:156043–156070, 2021.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. Gleu without tuning. *arXiv preprint arXiv:1605.02592*, 2016.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, pages 1–101, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Amarildo Rista and Arbana Kadriu. Automatic speech recognition: a comprehensive survey. *SEEU Review*, 15(2):86–112, 2020.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. mgpt: Few-shot learners go multilingual, 2022. URL https://arxiv.org/abs/2204.07580.

Annika Simonsen, Sandra Saxov Lamhauge, Iben Nyholm Debess, and Peter Juel Henrichsen. Creating a basic language resource kit for faroese. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, 2022. European Language Resources Association (ELRA). URL https://aclanthology.org/2022.lrec-1.495.pdf.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020.

Ferran Suay and Davyth Hicks. The european language equality network, proposals for better implementation of the ecrml. In *Minorities and Minority Languages in a Changing Europe*, Conference on the occasion of the 20 th anniversary of the Framework Convention for the Protection of National Minorities and the European Charter for Regional or Minority Languages, Strasbourg, France, June 18-19 2018. Council of Europe. URL https://rm.coe.int/speech-ferran-suay-elen/16808b754e.

José Manuel Ramírez Sánchez and Carmen García Mateo. Deliverable D1.15 Report on the Galician Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE__Deliverable_D1_15__Language_Report_Galician_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*, 2021.

Jörg Tiedemann. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. *arXiv preprint arXiv:2010.06354*, 2020.

Shamsa Umar, M.Shirin Maryam, Fizza Azhar, Sayyam Malik, and Ghulam Samdani. Sentiment analysis approaches and applications: A survey. *International Journal of Computer Applications*, 181:1–9, July 2018. doi: 10.5120/ijca2018916630.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537, 2019. URL http://arxiv.org/abs/1905.00537.

A. Appendix: The BLARK

BLARK MATRIX **100%**	Cross-cutting resources **40%**	Corpora **70%**
		NLP tools **10%**
		Lexical Resources **10%**
		Language Models **10%**
	LT tasks **60%**	Speech Synthesis **20%**
		Speech Recognition **20%**
		Machine Translation **20%**
		Other LT Tasks **30%**
		Benchmarking **10%**

Cross-cutting resources: corpora 70%			
Resource	**Size 30%**	**Quality 40%**	**License 30%**
Annotated corpora **20%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
	Size 40%	**Quality 30%**	**License 30%**
Reference corpus **30%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
Macro corpus **50%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	

Cross-cutting resources: lexical resources 10%			
Resource	**Size 70%**		**License 30%**
Dictionaries **50%**	Small **20%** Medium **60%** Large **100%**		Closed **15%** Restricted **30%** Open **100%**
	Size 30%	**Quality 40%**	**License 30%**
Annotated lexicons **50%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

Cross-cutting resources: NLP tools 10%			
Resource		**Existence 70%**	**License 30%**
POS tagging **20%**	Tokenization **5%**	Yes/No **100%**	Closed **15%** Research **30%** Open **100%**
Parsing **20%**	Lemmatization /word segmentation **5%**		
NERC **20%**	Word sense disambiguation **5%**		
Language identification **20%**	Coreference resolution **5%**		

Cross-cutting resources: language models 10%			
Resource	**Size 70%**		**License 30%**
Embeddings **10%**	**Vocabulary size 50%**	**Training corpus 50%**	
	Small **20%** Medium **60%** Large **100%**	Small **20%** Medium **60%** Large **100%**	Closed **15%** Research **30%** Open **100%**
	Nº Parameters 40%	**Training corpus 60%**	
Autoencoder (monolingual) **30%**	Small **20%** Medium **60%** Large **100%**	Small **20%** Medium **60%** Large **100%**	Closed **15%** Research **30%** Open **100%**
Autoregressive ´ (monolingual) **30%**	Small **20%** Medium **60%** Large **100%**	Small **20%** Medium **60%** Large **100%**	Closed **15%** Research **30%** Open **100%**
	Nº Parameters 40%	**% of the language in training data 60%**	
Autoencoder (multilingual) **15%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Research **30%** Open **100%**
Autoregressive (multilingual) **15%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Research **30%** Open **100%**

LT tasks: speech synthesis 20%			
Resource	**Size 40%**	**Quality 30%**	**License 30%**
Corpus **70%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
Model **30%**		**Quality 70%**	**License 30%**
		Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

LT tasks: speech recognition 20%				
Resource	**Size 40%**		**Quality 30%**	**License 30%**
	Nº hours 60%	**Nº speakers 40%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
Corpus **70%**	Small **20%** Medium **60%** Large **100%**	Small **20%** Medium **60%** Large **100%**		
Model **30%**			**Quality 70%**	**License 30%**
			Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

LT tasks: machine translation 20%				
Language pair	**Resource**	**Size 40%**	**Quality 30%**	**License 30%**
For a maximum of 4 pairs (=25% each)	Corpus **70%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
	Model **30%**		**Quality 70%**	**License 30%**
			Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

Other LT tasks: Grammatical Error Correction 20%			
Resource	**Size 40%**	**Quality 30%**	**License 30%**
Corpus **70%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
Model **30%**		**Quality 70%**	**License 30%**
		Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

Other LT tasks: summarization 20%			
Resource	**Size 40%**	**Quality 30%**	**License 30%**
Corpus **70%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
Model **30%**		**Quality 70%**	**License 30%**
		Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

Other LT tasks: sentiment analysis 20%			
Resource	**Size 40%**	**Quality 30%**	**License 30%**
Corpus **70%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
Model **30%**		**Quality 70%**	**License 30%**
		Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

Other LT tasks: fact checking 20%			
Resource	**Size 40%**	**Quality 30%**	**License 30%**
Corpus **70%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**
Model **30%**		**Quality 70%**	**License 30%**
		Low **20%** Medium **60%** High **100%**	Closed **15%** Restricted **30%** Open **100%**

Other LT tasks: dialog systems 20%				
Type	**Can you interact in your own language with...**			**Answers**
Usability **30%**	Mobile assistants **14,3%** LLM-based general chatbots and Q&A systems **14,3%** Smart speakers **14,3%** Public administration dialog systems **14,3%** Frequent e-commerce dialog systems **14,3%** Health applications dialog systems **14,3%** Other task-oriented dialog systems **14,3%**			Barely **20%** Partially **60%** Mostly **100%**
Corpora 70%	**Corpora size 20%**	**Corpora Quality 50%**		**License 30%**
Corpora for domain specific dialog systems **40%**	Small: **20%** Medium: **60%** Large: **100%**	**Annotations 50%**	**Number of domains 50%**	Closed **15%** Restricted **30%** Open **100%**
		Low **20%** Medium **60%** High **100%**	Few domains **20%** Some domains **60%** Many domains **100%**	
	Corpora size 40%	**Corpora Quality 30%**		**License 30%**
Corpora for general generative dialog systems **60%**	Small **20%** Medium **60%** Large **100%**	Low **20%** Medium **60%** High **100%**		Closed **15%** Restricted **30%** Open **100%**

Benchmarking 10%		
Taking into account the LT tasks mentioned above...		
Resource	**Question**	**Answer**
Widely used evaluation resources **60%**	Is your language present in mainstream benchmarks or evaluation resources/datasets?	Barely **20%** Partially **60%** Mostly **100%**
Evaluation materials of its own **40%**	Does your language have sufficient evaluating resources of its own?	

B. Appendix: The BLARK for Galician (GL-BLARK)

GL-BLARK 54,03% /100%	Cross-cutting resources 24,91%/40%	Corpora 40,28%/70%
		NLP tools 9%/10%
		Lexical resources 7,15%/10%
		Language models 5,86%/10%
	LT tasks 29,12%/60%	Speech synthesis 20%/20%
		Speech recognition 16,08%/20%
		Machine translation 8,32%/20%
		Other LT tasks 2,14%/30%
		Benchmarking 2%/10%

Cross-cutting resources: corpora 40,28%/70%			
Resource	**Size**	**Quality**	**License**
Annotated corpora 13,6%/20%	High **30%**	Low **8%**	Open **30%**
	Size	**Quality**	**License**
Reference corpus 13,95%/30%	Medium **24%**	Medium **18%**	Closed **4,5%**
Macro corpus 30%/50%	Medium **24%**	Low **6%**	Open **30%**

Cross-cutting resources: lexical resources 7,15%/10%			
Resource		**Size**	**License**
Dictionaries 37,25%/50%		Large **70%**	Closed **4,5%**
	Size	**Quality**	**License**
Annotated lexicons 34%/50%	Large **30%**	Low **8%**	Open **30%**

Cross-cutting resources: NLP tools 9%/10%	
Existent Resource	**License**
POS tagging 20%/20%	Open **30%**
Parsing 20%/20%	
NERC 20%/20%	
Language identification 20%/20%	
Tokenization 5%/5%	
Lemmatization/word segmentation 5%/5%	

Cross-cutting resources: language models 5,86%/10%			
Resource	**Size**		**License**
Embeddings 5,8%/10%	**Vocabulary size**	**Training corpus**	
	Medium **30%**	Small **10%**	Open **30%**
	Nº Parameters	**Training corpus**	
Autoencoder (monolingual) 35,52%/40%	Medium **24%**	Large **60%**	Open **30%**
Autoregressive (monolingual) 0%/20%			
	Nº Parameters	**% of the language in training data**	**License**
Autoencoder (multilingual) 13,28%/20%	Large **40%**	Small **12%**	Open **30%**
Autoregressive (multilingual) 4,09%/10%	Large **40%**	Small **12%**	Closed **4,5%**

LT tasks: speech synthesis 20%/20%			
Resource	**Size**	**Quality**	**License**
Corpus 70%/70%	Large **40%**	High **30%**	Open **30%**
Model 30%/30%		**Quality**	**License**
		High **70%**	Open **30%**

LT tasks: speech recognition 16,08%/20%			
Resource	**Size**	**Quality**	**License**
Corpus 50,4%/70%	Medium **24%**	Medium **18%**	Open **30%**
Model 30%/30%		**Quality**	**License**
		High **70%**	Open **30%**

LT tasks: machine translation PAIR 1: ES-GL 20,8%/100%			
Resource	**Size**	**Quality**	**License**
Corpus 61,60%/70%	Large **40%**	Medium **18%**	Open **30%**
Model 21,60%/30%		**Quality**	**License**
		Medium **42%**	Open **30%**

LT tasks: machine translation PAIR 1: EN-GL 20,8%/100%			
Resource	**Size**	**Quality**	**License**
Corpus 61,60%/70%	Large **40%**	Medium **18%**	Open **30%**
Model 21,60%/30%		**Quality**	**License**
		Medium **42%**	Open **30%**

LT tasks: machine translation global result 8,32%/20%		
ES-GL	**EN-GL**	**TOTAL**
20,8%	20,8%	41,6%

Other LT tasks: Grammatical Error Correction 2,59%/20%			
Resource	**Size**	**Quality**	**License**
Corpus 12,95%/70%	Small **8%**	Low **6%**	Close **4,5%**
Model 0%/30%			

Other LT tasks: summarization 0%/20%	
Corpus 0%/70%	0%
Model 0%/30%	

Other LT tasks: Sentiment Analysis 0%/20%	
Corpus 0%/70%	0%
Model 0%/30%	

Other LT tasks: Fact Checking 0%/20%	
Corpus 0%/70%	0%
Model 0%/30%	

Other LT tasks: Dialog systems 4,55%/20%			
Type	**Can you interact in your own language with...**		
Usability 4,29%/30%	Mobile assistants 0%/14,3% LLM-based general chatbots and Q&A systems 2,86%/14,3% Smart speakers 0%/14,3% Public administration dialog systems 2,86%/14,3% Frequent e-commerce dialog systems 2,86%/14,3% Health applications dialog systems 2,86%/14,3% Other task-oriented dialog systems 2,86%/14,3%		
Corpus 18,48%/70%			
Corpora for domain specific dialog systems 0%/40%			
	Corpora size	**Corpora Quality**	**License**
Corpora for general generative dialog systems 26,4%/60%	Small **8%**	Low **6%**	Open **30%**

LT tasks: Other LT tasks 2,14%/30%	
TASKS	**RESULTS**
GEC	2,59%/20%
Summarization	0%/20%
Sentiment Analysis	0% /20%
Fact Checking	0%/20%
Dialog Systems	4,55%/20%

LT tasks: Benchmarking/quality evaluation 2%/10% Taking into account the LT tasks mentioned above...		
Resource	**Question**	**Answer**
Widely used evaluation resources 12%/60%	Is your language present in mainstream benchmarking datasets?	Barely
Evaluation materials of its own 8%/40%	Does your language have evaluating resources of its own? (gold-standards, adversarial datasets, etc.)	Barely