Chapter 9

Inclusive Host Language Teaching:

Official Texts for Migrant and Refugee People

Raquel Amaro
https://orcid.org/0000-0002-4923-7186
NOVA University of Lisbon, Portugal

Susana Correia
NOVA University of Lisbon, Portugal

Matilde Gonçalves
NOVA University of Lisbon, Portugal

Chiara Barbero
NOVA University of Lisbon, Portugal

Miguel Magalhães
https://orcid.org/0000-0003-0055-8971
NOVA University of Lisbon, Portugal

ABSTRACT

This chapter presents research on the teaching-learning of Portuguese as a host language, based on the exploration of authentic informational and institutional texts targeting migrant and refugee people, and considering that successful host language teaching must correspond to the needs of its target audience. The chapter discusses methods of defining and identifying criteria and features to monitor official texts with regard to inclusiveness and bias. It provides insights on how to select real texts to be used in task-based language teaching approaches for inclusive host language teaching. Departing from a real corpus analysis, the potential and the limitations of existing guidelines to inclusiveness for the assessment of real texts are shown, as well as other still neglected issues. Furthermore, this chapter provides future research directions to an effective and reliable assessment of inclusive texts that can serve as inclusive host language teaching materials through NLP and machine learning approaches.

DOI: 10.4018/978-1-7998-8579-5.ch009

INTRODUCTION

Gathering knowledge and expertise from several subareas of Linguistics, the work discussed in this chapter concerns preliminary research on the teaching-learning of Portuguese as a host language through Natural Language Processing (NLP) approaches to real texts, aiming at informing machine learning systems. Successful host language teaching must correspond to the needs of its target audience – migrant population and people on the move for the widest range of reasons. Also, the needs of the hosting community must be met, considering, in particular, the goals of a well-succeeded and inclusive integration for recently arrived people. This means providing relevant and pragmatic information regarding many aspects of daily life, on the one hand, and of institutional and legal constraints, on the other hand. These two aspects should enable an autonomous, free and unconstrained life and participation in society for those arriving in a host-country.

Language skills have been considered as a form of host-country specific human capital in Economics since the early 1980s (Carliner, 1981; McManus et al., 1983), and several studies clearly establish the negative consequences of language barriers in fields such as health care (Jaeger et al., 2019), labour market performance and income (Chiswick & Miller, 2014; Miranda & Zhu, 2013), demographic outcomes, children education and residential choice (Bleakley & Chin, 2010; Chen, 2013; Guven & Islam, 2015). Language learning is thus a key factor that promotes socio-economic integration, as well as access to rights and services (CoE, 2018).

Task-Based Language Teaching framework may contribute directly to the achievement of these goals by providing the means to train use-oriented language skills in an integrated way, thus being particularly adequate for inclusive host language teaching (Ellis, 2017; Long, 2014).

Having these two quite simple assumptions in mind, this chapter has as main objectives:

- To discuss methods and ways of defining and identifying criteria and features to monitor official (informational and institutional) texts targeting migrant and refugee people, with regard to inclusiveness and bias,
- To provide insights on how to select real texts to be used in task-based language teaching approaches for inclusive host language teaching.

It is organized in five main sections presenting: (1) the background and related work concerning the concepts and approaches to integration in general, and integration through language in particular; (2) Portuguese official texts targeting migrant and refugee people, with regard to the research issues pursued and methods used, and the set of data collected; (3) the linguistic cues and constructions expressing bias or inclusiveness, namely in what concerns the type of information extracted from the corpus and the results obtained; (4) the assessment and selection of materials for inclusive host language teaching, accounting for semantic values relevant for inclusiveness, and recommendations for selecting appropriate language teaching materials; and (5) final remarks and future research directions on featuring inclusiveness in texts.

BACKGROUND AND RELATED WORK

Integration and Inclusion

As far as the issue of reception, inclusion or integration of people is concerned – encompassing but not restricted to people with disabilities, children and young people, migrant people, among others -, it is interesting to observe the use and assumptions of these specific concepts in order to understand how they can affect social policies. Being host language teaching regarded as an essential aspect of integration and/or inclusion of people on the move, it is a relevant part of these social policies. Looking at the use of these concepts over time, it can be understood that there has been a linguistic change favoring the resource of *inclusion* in detriment of *integration*, as advocated by the Salamanca Statement for action on special needs education (UNESCO, 1994).

In terms of official documents or institutions, as we can observe below, both expressions are used without differentiation or even the preference of one of them over the other:

The European way of life is an inclusive one. Integration *and* inclusion *are key for people coming to Europe, for local communities, and for the long-term well-being of our societies and the stability of our economies. If we want to help our societies and economies thrive, we need to support everyone who is part of society, with* integration *being both a right and a duty for all. (CE, 2020, p. 2)*

Portugal has registered a very positive evolution since 2003, both in terms of policies and practices of immigrant reception and integration*, to which not only the interventions of the State at national level, but also that of local authorities, civil society organisations and immigrant communities themselves have contributed. (ACM, n.d.)*

However, *integration* and *inclusion* may have different meanings, with impact on the conception and representation of the people and institutions that receive and of the people who are received. According to ECRI – European Commission against Racism and Intolerance of the Council of Europe, integration is a process on both sides of society and public authorities towards people striving for it. Inclusion, on the other hand, is a process of accommodation, conciliation, and adaptation that values diversity, sees it as an asset, and aims at creating the conditions for the active participation of all members of society (ECRI, n.d.)

Looking at the most recent official documents on the integration and inclusion of migrant people, the Action Plan on Integration and Inclusion 2021-2027 (CE, 2020), it is possible to notice that the word *integration* was used 186 times in the document while *inclusion* was used only 71 times. Although the terms *integration* and *inclusion* appear frequently, and probably most of the time together, a differentiation in terms can be observed. Integration and inclusion can and should be a win-win process, benefiting the entire society. But if integration and inclusion are to be successful, it must also be a two-way process whereby migrants and EU citizens with migrant backgrounds are offered help to integrate and they in turn make an active effort to become integrated. These differences are clearly stated in the Action Plan on Integration and Inclusion 2021-2027 (CE, 2020), as presented here below:

The integration *process concept involves the host society, which should create the opportunities for the immigrant people full economic, social, cultural, and political participation. It also involves adaptation* *by the migrant people who are supposed to have rights but also responsibilities in relation to their new country of residence. (CE, 2020, pp.1-2).*

Inclusion *(for all) is about ensuring that all policies are accessible to, and work for, everyone, including migrant people and EU citizens with migrant background. This means adapting and transforming mainstream policies to the needs of a diverse society, taking into account the specific challenges and needs of all the groups involved. Actions to help migrant people integrate need not, and should not, be at the expense of measures to benefit other vulnerable or disadvantaged groups or potential minorities. On the contrary, these actions contribute to making overall policies more inclusive. (CE, 2020, p. 5).*

The relation between these concepts of integration and inclusion and the use of language is discussed in the next section.

Integration and Inclusion through Language

All action and policies for integrating and including migrant persons depend on language as mediation of communication. Learning the language of the host country can be a crucial step in order to successfully integrate. However, the process should not either be focused on the recently arrived, nor stop a few months after arrival. Language teaching and learning, including the access to more formal language teaching classes, should be continuously supported, also for intermediate and advanced courses, and tailored to the needs of different groups, as it is, for instance, when it targets investors or highly technical or specialized human resources such as medical personnel.

Combining language training with the development of other skills or work experience and with measures like childcare has proven to be particularly effective in improving access to and the outcome of language training. Gaining an understanding of the laws, culture, and values of the society they are part of as of arrival as early as possible, for example through civic orientation courses, is crucial for migrant people to fully participate in their new community. (CE, 2020, p. 10)

Linguistic Bias, Linguistic Aggregation, and Host Language

As the Council of Europe highlights, languages (and language in general) play a key role in people's life and in the functioning of society as a democratic system. (cf. The Language Policy Programme of the Council of Europe[1]). In this sense, languages assume a role of mediation between different parts of society, contributing in an indispensable way to the integration and inclusion of migrant people. In this process of integration and inclusion, it is clear that the ultimate goal is for the migrant person to become a full citizen, with equal rights and duties, and actually belonging to a community "which entrains politics and rights, notably political rights." (CDCC, 2000, p. 16). Citizenship and individual rights are necessarily related to non-discrimination, individual (informed) freedom and to citizen participation in the definition of governments and laws. Consequently, information contained in legislative texts concerning immigration matters, informative texts for job and housing applications, tax declaration, healthcare insurance, among others must be accessible, comprehensible, and inclusive in order to facilitate the integration process of migrant persons, avoiding discriminatory representations.

The intertwined relationship between language and society has long been pursued by linguists, psychologists, and sociologists (for a review, see Coulmas, 2017). Depending on the perspective taken, different contextual aspects of language use – whether social, cultural, or political – can be considered. Starting from a conception of language as a system of relationships between linguistic material and the context of communication, the angle of analysis may be based on the textual producer and the use he/she makes of it in relation to the communication situation (lexicon and forms of address, for example).

It is from this (tri)angle – communication context/textual producer/language use – that it is possible to detect and analyse the social and individual representations that one has of oneself, the others, and the society. In fact, the use of language is never neutral (Del Valle, 2007); (Voloshinov, 1929/2014), since it is always a product of social interaction, expression of ideological positions, power, authority or legitimacy (Foucault, 1970, 1972). Studies on linguistic bias demonstrate how words reflect social influence and the maintenance of stereotypes (Beukeboom & Burgers, 2017), (Borchmann et al., 2019). The existence of these studies raises the current need to understand how biases shape our perception, judgements and interaction with the world and with human beings, and how they create discrimination and a non-cohesive society.

In addition to the study on biases dealing with aspects at a more specific level, it is important to mention works on ideology, namely, Fairclough (1989), Van Dijk (2014) and Voloshinov (1929). In fact, this notion is equivalent to a set of socially shared beliefs, circulating at a more global level, that are created and embedded in language, our commonest form of social behavior. The study of unconscious biases in human behavior made us aware that, up to a certain point, we cannot escape our societal experiences.

Language being a human and social product, it is also true that it can play a major role in shifting social attitudes (Caliskan et al., 2017; De Houwer, 2006; Mann et al., 2020). Both biases and ideologies can be negative, conveying a set of harmful and discriminative representations, or, preferably, they can have positive outcomes, transmitting and building a system of beliefs that is aggregative and beneficial to society. Texts, as a global communication activity and a product of human interaction in a communication situation, are the place where social, cultural, historical, and linguistic aspects intersect (Adam, 2005; Bronckart, 1999; Rastier, 2001). Consequently, texts are where the creation and manifestation of social representations, and consequently of biases, can be observed.

In the case of the situation of reception and integration of migrant people, the issue of preventing discrimination is central. Thus, producing texts or teaching material that, in addition to providing all the necessary information for adaptation and community integration and participation, truly respects the person's identity and particularities is revealed as truly urgent in the current context.

Task-Based Host Language Teaching

The programs, activities, and materials available for language teaching and learning are often too broad and general in their purposes, not covering or having specific concerns directly related to specific target audiences or purposes. For instance, vocabulary areas such as family, house, leisure, or food are usually considered in general language teaching curricula. However, if we consider the needs of migrant and refugee language learners, these topics are clearly inadequate or irrelevant, especially in urgent or severely vulnerable situations (Elsod & Marques, 2019). Therefore, developing teaching materials that train migrant people on real-world tasks involving inclusive language is an essential step for early and successful integration and effective inclusion. According to specific surveys and already existing studies, such tasks should include specific real-life situations such as requesting social protection, making a doctor appointment, opening a bank account, going to a pharmacy, or reading and negotiating a rent contract (Elsod & Marques, 2019; OCDE, 2018).

Task-Based Language Teaching (TBLT) offers the ideal theoretical and applied tools to boost integration through language. TBLT is an evidence-based approach to language teaching, curriculum design and materials development that has received a substantial amount of empirical support from second language research over the past two decades (Ellis, 2017; Long, 2014). From the aforementioned studies, among others, there is a clear consensus that task-based learning promotes second language learning, and that this is an educationally effective manner of teaching a novel language (Long, 2014). Communicative tasks used in the TBLT approach are sequenced from simple to complex, following research on the effects of task complexity on language learning and defined according to the learners' proficiency levels (Robinson, 2011). In this framework, tasks are language-learning activities that satisfy the properties below:

1. There is a real-world relationship, i.e., tasks address the actual communicative needs of learners.
2. The primary focus of the task is on meaning and not on linguistic form.
3. There is a communication "gap" that should be addressed by the successful completion of the task.
4. There is a clearly defined outcome other than the use of language, i.e. the language serves as the means for achieving the outcome, it is not an end in its own right.

The analysis of informative texts that target migrant people produced by hosting institutions can be a helpful resource to develop and deploy classroom materials that are useful for teaching a host language, by facilitating the knowledge of migrant people's specific needs and promoting a more dedicated approach to language pedagogy and didactics. It is a necessary step to assure that the materials used in language learning/teaching tasks express inclusive attitudes, thus adding to the motivation for language study, on the one hand, and to the successful integration/inclusion of migrant people, on the other. However, as discussed in the remainder of the chapter, it relies heavily on a thorough, informed, and objective analysis of real texts, a time consuming and demanding task. NLP tools and automated approaches can provide a means to facilitate this endeavor.

Assessing Features and Trends through NLP and Machine Learning Approaches to Texts

NLP methods and tools are widely used to perform all kinds of analysis on language data, in particular machine-learning systems. The great advantage of these approaches is that they allow for the systematic and immediate analysis of large quantities of data at low human-resources cost. Methodologically, the systems are designed and tuned considering three essential aspects: data, machine-learning process, and result evaluation/comparison.

Machine learning consists in the processing strategy to infer patterns or functions from data. In supervised machine learning systems, a sample of the data and of the expected results is provided, i.e., the system has access to categorized/annotated data and must devise the best function to approximate the input data to the output data, based on the reference data/expected results. In an unsupervised machine learning system, there is no baseline data or expected results; the system must infer a natural present structure in unannotated data. In both cases, the input data is crucial since the systems apply algorithms and statistical models, with or without explicit instructions, but with power of inference, to determine what is the best function to arrive at the expected results from patterns/traits inferred and or extracted from the input data. Depending on the methods used, the data needed to develop the systems may differ.

Machine learning systems do not have explicit linguistic knowledge rules, but use implicit linguistic knowledge, reflected in the patterns of the language data, including bias.

Several studies have shown a growing concern for the mitigation of bias in the construction of models in NLP. Studies such as those by Bolukbasi (2016) showed how the bias is present and has been perpetuated in the data used in NLP if there are no mitigation strategies. Focusing on gender stereotypes, Bolukbasi (2016) shows how word embedding, trained only with co-occurrence in corpora, can convey a sexist interpretation, and trivial counting of words and word co-occurrence can be misleading:

For instance, the term male nurse is several times more frequent than female nurse (similarly female quarterback is many times more frequent than male quarterback). (Bolukbasi, 2016, p. 3)

This shows how the word *nurse*, for example, has associated a gender stereotype in which the general rule is to associate the word with the female gender and when it is associated with the male gender it co-occurs with the word *male*. This is just one of the examples of how co-occurrence can reveal and perpetuate a stereotype, in this case associated with a profession.

Regarding stereotypes associated with migrant people, there are no specific studies on bias in NLP. However, it is important to emphasize that the problem of stereotypes associated with migrant individuals has been discussed from a linguistic and social perspective (cf. Rosa, 2019). Other studies have been developed in various areas such as the racial bias Avineri et al. (2019) and Blodgett (2021). The latter, having the African-American language as its object of study, adopts an intersectional analysis that brings together perspectives from sociology, sociolinguistics, social psychology, among others, and proposes a taxonomy of the harm that emerges from the use of word-based NLP systems embedding. According to this work, NLP systems are democratized and part of everyday life, but the genesis of NLP is, in itself, unfair and biased since it is based on models, generalizations and stereotypes, and since the data that serve as a model may reflect systemic problems. Analysis of stereotyping are usually based on word embeddings analyses and coreference resolution systems and should be further grounded in a solid “understanding of the empirical reality of unjust social arrangements.” (Blodgett, 2021, p. 98).

One of the problems that arises with these “empirical reality of unjust social arrangements” is that these phenomena are analyzed by the social sciences through a qualitative analysis, since they are theoretical variables, and do not have a measurable cut. The solution for this problem involves the identification and observation of features or properties that may be representative for the variable under analysis, inclusiveness and bias in institutional texts targeting migrant and refugee people, and the definition of a measurement model and the development of methodological tools to assess it.

PORTUGUESE OFFICIAL TEXTS TARGETING MIGRANT AND REFUGEE PEOPLE

Issue and Methodology

Communicative situations in Portuguese as a Foreign Language (PFL) teaching unfortunately remain an obstacle to integration and inclusion of people arriving in the country. Relevant language use is insufficient or unapproachable to migrant learners, in particular to those in vulnerable situations, aiming at achieving autonomous participation in the community as soon as possible. Current PFL materials still focus on inadequate or irrelevant thematic contents (e.g., family, leisure, food, or shopping). They disregard the specificities of integration and inclusion processes (e.g., looking for a home, finding childcare, registering in a local healthcare facility, accessing the job market) and, ultimately, they promote individual and cultural segregation.

The present study addresses the research question of how institutional and informative texts can provide linguistic cues to help design teaching materials that can contribute to an inclusive host language teaching. The hypothesis put forth is that the information material targeting migrant people is (i) heavily constrained by the use of lexical and grammatical constructions that convey implicit biases towards non-inclusion and discrimination, while (ii) reflecting neutrality concerns that make texts impersonal and non-inclusive and aggregational.

To pursue a solid TBLT approach to an inclusive host language teaching, it is necessary to first assess texts in terms of their inclusiveness and aggregational properties. To this end, i.e., to analyze official texts targeting migrant and refugee people, produced by institutions with special interest and relevant action in the integration and inclusion of people coming to Portugal, with regard to their inclusiveness and aggregational properties, we devised the following method:

Step 1: Collection of relevant materials
Step 2: Review of existing studies and guidelines
Step 3: Application of NLP tools to the data
Step 4: Determination of assessment measures related to inclusiveness
Step 5: Analysis and systematization of results.

The **collection of relevant materials** phase concerns the determination of the relevant universe, considering the Portuguese current reality and the processes and information that are useful and necessary to the people these institutions aimed to work with. This determination will afterwards inform the corpus selection and representativeness criteria. This step will result in the analysis corpus.

The **review of existing studies and guidelines** concerning inclusive language is necessary to list the communication principles at stake and the specific linguistic expressions that are related to these principles. It requires the systematization of already established guidelines and their adaptation to the topic of integrating people in Portuguese society, on the one hand, and to the linguistic specificities of the European Portuguese language, on the other. In this phase, a list of Portuguese expressions and constructions relevant for inclusiveness assessment will be compiled.

The **application of NLP tools to analyze the data** concerns the work on corpus shallow processing and exploitation using existent tools. In this phase, we used Sketch Engine to host the corpus and shallow process it, adding it Parts-of-Speech (POS) annotation and lemmatization. This specific tool allows for extracting word lists, collocates and concordances, associated with sophisticated frequency and statistical information. It also provides a Corpus Query Language (CQL) query system that allows for the search of complex expressions – purely structural, using only POS information, purely lexical, using string with or without POS information, and hybrid, using both lexical and morphosyntactic information. This stage results in the extraction of word lists, expressions and statistical information from the compiled corpus.

The **determination of assessment measures related to inclusiveness** phase of the work encompasses the analysis of the data extracted from the corpus, to identify linguistic features and establish objective relations between the observed salient or relevant features and integration/inclusion aspects or goals that can allows us to characterize a given text as more or less inclusive. It comprises quantitative analysis of the data, informed by the NLP tools and methods used and by the validation/correlation of the results achieved regarding the guidelines established for inclusive language. This analysis is necessarily balanced by qualitative analysis to assure that the aspects under inspection are actually the ones expressed or reflected in the data.

The **analysis and systematization of results** stage consists in the final stage of the work to identify the linguistic features patterns and generalizations (for instance, in terms of presence or absence, or distribution with regard to the type of texts occurring in the corpus) that relate to inclusion/integration, but also to pinpoint gaps in the methods or in already established knowledge, able to motivate further work and research directions.

The activities and results of these different phases are further discussed in the next sections of this chapter.

MIGRANTE.PT Corpus

MIGRANTE.PT is a monolingual corpus (European Portuguese), oriented for specific purposes. It encompasses around 1.5 million tokens of institutional texts concerning topics relevant for the integration/inclusion of migrant people in Portugal, and collects authentic texts freely available online from the most relevant Portuguese institutions, responsible for managing and assisting in the integration of migrant and refugee people – ACM Portugal: The High Commissariat for Migrations, Portuguese Refugees Council –, governmental and non-governmental institutions such as Caritas, National Entity for Health Regulation, Migrations Observatory; Foreign and Borders Services; Ministry of Internal Administration; Social Security; Portuguese Parliament and several municipalities; as well as media texts from several sources, collected by these organizations according to their informative and/or promotional character.

As a language for specific purposes corpus, i.e., a repository of real linguistic data framing a very specific communicative context, MIGRANTE.PT offers interesting insights about language users' needs, on the one hand, and the language learners need to acquire in order to successfully engage in social practices, on the other (Cotos, 2017).
In this particular case, the communicative context covers:

- **Heterogeneous Participants**: institutions or associative organization (source) and adult people involved in migration and accommodation processes (target), with or without the official and legal status of migrant or refugee individuals, with different levels of education, coming from diversified social and economic situations and having different linguistics backgrounds,
- **Official Texts**, that constitute the main communicative channel between institutions and the target public (as described in section *Corpus constitution and compilation* and
- **Specialized Lexicon**, encompassing all the aspects related to integration/inclusion in a host country.

Since corpus compilation always presupposes the act of sampling, it is crucial to guarantee the representativeness of the sample collected with regard to the universe of analysis that is intended to be represented and reflected in the corpus.

Corpus Design and Representativeness

Representativeness takes into account several aspects, linguistics and situational/extra linguistics, along with the specific goals the corpus intends to serve. For instance, strict statistical representativeness – that benefits the more frequent cases within the sampling universe, i.e., the sample is a direct extrapolation of the universe occurrences statistics – can be replaced by diversity representativeness – that privileges the insertion of all the types of occurrences, regardless of frequency of occurrence, i.e., the sample covers all the types of cases, even if infrequent. For this reason, representativeness is only partially measurable in purely quantitative terms. Representativeness assessment may not rely on consensual and universal parameters (Biber, 1993; McEnery & Hardie, 2012) and often depends on particular and more subjective aspects.

In this way, the intrinsic quality, relevance, and representativeness of the MIGRANTE.PT corpus with regard to the intended research goals are supported by the reliability of the sources considered, which cover all the relevant institutions associated to migration management, hosting and integration/ inclusion of migrant people in Portugal. These consider, thus,

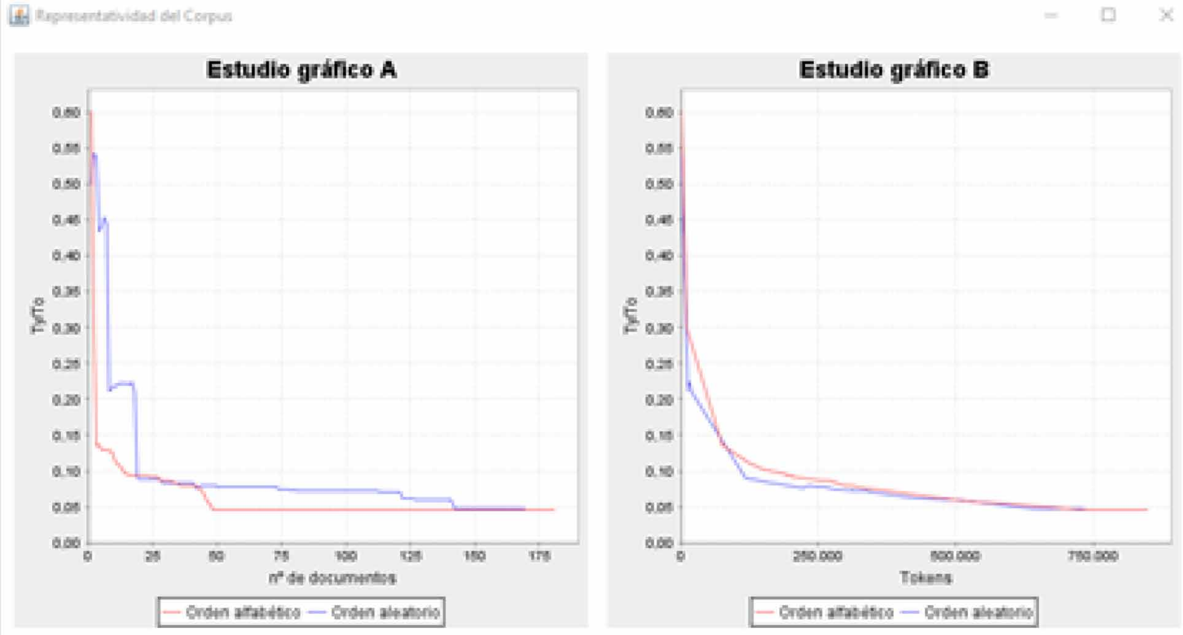
- **The Existing Institutions**: governmental (ex.: Migrations Observatory) and non-governmental (ex.: Portuguese Refugees Council),
- **The Different Ranges of Territorial Coverage and Actions**: national institutions (ex.: The High Commissariat for Migrations), regional institutions (ex.: Caritas), and local institutions (ex.: municipalities),
- **The Different Areas Associated to the Integration/Inclusion Process**: legal (ex.: Foreign and Borders Services; Ministry of Internal Administration), social services (ex.: Ministry of Internal Administration; Social Security), health (ex.: National Entity for Health Regulation), education (ex.: Municipalities), employment (ex.: Social Security, Portuguese Refugees Council).

In terms of size, the corpus includes all the online digital texts convertible to .txt format targeting migrant and refugee people. available on the webpages of these institutions in November 2020.

The application of specific statistical measures to the corpus, in particular the ReCor 2.3, ensures a satisfactory degree of representativeness by a statistical evaluation in terms of language sampling by means of lexical saturation assessment (see Figure 1).

Both in graphics A and B, the vertical axis represents the type/token ratio, while the horizontal axis shows the number of texts included in the corpus, in graph A, and the number of tokens, in graph B. In both steps of analysis, the graph demonstrates a substantial decrease of the type/token ratio as texts (in A) and tokens (in B) increase. Red lines represent the increase/decrease tendences considering files alphabetically ordered and the blue lines represent the same tendency considering randomly ordered files. This double operation is carried out to ensure that the order of the texts has no relevant repercussions on the results. Following Corpas-Pastor & Seghiri (2010), we can state that that the corpus is representative, at the point where both the red and blue lines stabilize. The MIGRANTE.PT corpus is representative around text 150 (graph A) and around at some point between 500.000 and 750.000 tokens. In total, MIGRANTE.PT has 1.5 million tokens.

Figure 1. Recor quantitative report on the lexical saturation of MIGRANTE.PT



Corpus Constitution and Compilation

The texts were collected from institutional web pages using BootCaT Frontend 1.31[2], a tool that simplifies the process of collecting data in text format from the www. BootCaT automates the process of collating texts in a single corpus, allowing various levels of control. The list of relevant URLs was collected according to the criteria already described and provided to the tool. The actual web pages are then retrieved, converted to plain text and saved in .txt format. A copy of the original files, as well as the complete URL of the retrieved file, are also saved, providing a solid backup of the data. The corpus is thus ready to be used by most concordancers. The corpus was then uploaded into SketchEngine[3] where it was compiled and shallow processed (POS tagging and lemmatization) for Portuguese[4].

In terms of its internal constitution, MIGRANTE.PT corpus gathers several types of texts, organized according to the topic and the purpose.
In particular:

- Informative texts (e.g., job and housing applications, tax declaration, healthcare insurance, school registration forms, citizenship request)
- Legal texts (e.g., Portuguese legislation concerning immigration matters),
- Dissemination texts (e.g., news and media).
- Technical texts (e.g., reports and plans on local and/or national projects and initiatives and guidelines)

The media texts were collected via these organizations, from several sources, and reflect the perception of these host institutions of events that can help integration or that report relevant situations.

Table 1. MIGRANTE.PT corpus constitution

Genres of Text	Number of tokens
Informative texts	122 651
Legal texts	143 236
Dissemination texts	90 981
Technical texts	1 078 683
Total	1 435 551

MIGRANTE.PT is accessible at the webpage of the Linguistics Research Center of NOVA University of Lisbon, freely available for download under Creative Commons conditions for non-commercial purposes. The resource includes linguistic data archives, i.e., texts in txt format, for download or shared through Sketch Engine, and the documents metadata in .xlsx format, including the follow information: (i) file name, (ii) author, (iii) text type, (iv) date, (vi) URL.

LINGUISTIC CUES AND CONSTRUCTIONS EXPRESSING BIAS OR INCLUSIVENESS

To inform the next step on the actual collection and analysis of MIGRANTE.PT corpus data, we performed a thorough review of existing studies and guidelines, covering both general indications from international institutions, language specific instructions for Portuguese (including both Brazilian and European Portuguese) and national guidelines[5] focusing on Portuguese institutions. Inclusive language guidelines (e.g., OHSU, 2021; OCDE, 2021; CES, 2021) usually refer to four major principles, which are not independent or disassociated from each other, as briefly summarized and briefly exemplified below.

1. **People First**: using formulations that reflect that people are not defined by their (permanent or temporary) conditions (origin status, disability, etc.), shifting the focus from the characteristics to the individual, as ‘person with/has deafness’ instead of ‘deaf’, ‘person with obesity’ instead of ‘obese’, ‘people on the move’ rather than ‘migrants’, ‘person with a substance use disorder’ rather than ‘addicts’, “individuals with physical/cognitive disability or disorder’ rather than ‘disabled’, ‘people who are homeless’ rather than ‘homeless’, etc.
2. **Specificity** instead of generalization: the use of vague expressions and generalizations can lead to unnecessary and uncomfortable discrimination. For instance, ‘diverse’ as a way to refer to people that are not white/black/Asian/national or that do not fit a given ‘norm’ (ex.: heterosexual) immediately creates a barrier between you and people who are different from you/the norm. *Strictu sensu*, ‘diverse’ denotes two or more people that are not equal. “Self-identification, as well as the labels others choose, can impact individuals’ integration, including feelings of belonging, attitudes and experiences.” (Portes & Rumbaut, 2001).
3. **Tolerant/positive perspective** instead of judgmental: the use of assertive, and usually less tolerant, expressions to describe specific circumstances or states can express strict judgments, even when these are not intended. For instance, labeling a patient as ‘noncompliant’ can express the presup-

position that the patient did not want to take the medication willingly. The same can be expressed more neutrally by expressions such as 'did not complete treatment' or 'stopped taking medication'.

4. **Gender neutrality**: expressions specifying one gender – usually masculine – to refer to groups including other genders help promote gender inequalities and underrepresentation of groups. In this case, use 'all' instead of 'ladies and gentlemen', 'pregnant people' instead of 'pregnant women' or 'parents' instead of 'mother(s) and father(s)'. Special attention to names of professions is also cautioned, as in the example 'cleaning personnel' instead of 'cleaning ladies'.

Considering the specific case in study here, namely the situation of people on the move, including migrant and refugee people, the aspect of **condescension/patronizing attitude** was also referred to. In this case, many times the intention to characterize and emphasize the harsh situations of a given vulnerable group can lead to the non-intended representation and perception of deeply distressed and non-autonomously functional individuals that require a patron or sponsor figure. This can be reflected in texts by the overuse of expressions such as 'victim', 'aid', 'needed', etc.

The analysis of existing studies informed a first data extraction plan. In particular, to assess non-inclusive uses, we started by listing specific relevant words, and Portuguese equivalents, stated in these documents and according to these aspects:

- People first and tolerant perspective: *migrante, estrangeiro, refugiado, requerente de asilo, imigrante*
- Specificity: *norma, normal, típico, standard, diverso, diferente, verdadeiro, natural, especialmente, outros, maioria, maioritário, regular, nativo, nacional*
- Gender neutrality: *mulher, homem, pai, mãe, marido, rapaz, rapariga, menino, menina*
- Condescension/patronizing attitude: diminutives (*-inho*), *auxílio, ajuda, vítima, necessidade, precisar, carente, carenciado, necessitado, corajoso, superar, superação, minoria, conflito, sofrimento, sofrer, guerra*.

A similar process was also done regarding inclusiveness, i.e., expressions that can indicate the use of inclusive and aggregational language. From the analysis of these guides, we collected the following expressions: *pessoa, indivíduo, cidadão, utente*, *pessoa/indivíduo com* (person, individual, user, person/individual with) group nouns such as *juventude*, *comunidade*, *classe*, *grupo* (young people, community, class, group), 1st person personal pronouns and possessive (*eu, nós, meu, minha, nosso* (I, us, mine, our) and verb forms in the 1st person.

To further populate these lists with other relevant examples, we verified the lists of frequent nouns, adjectives, and verbs, typical to this specific type of texts, as detailed in the next subsection.

Extracting Information from the Corpus

The initial analysis to look for candidates was performed on the tagged and lemmatized MIGRANTE.PT corpus, using Sketch Engine and according to a strict method to identify the specific lexicon of the domain – integration/inclusion of migrant people. So, lists of nouns, adjectives, and verbs, with information on the specific frequency of each word, were extracted. These frequencies were then balanced, i.e., absolute frequencies were converted into relative ones according to the total of tokens (running words) of the corpus using the formula

Relative Frequency=Absolute frequency/Total of tokens of the corpus.

Lists of relevant items of each POS list from MIGRANT.PT were then collected and integrated with lists of a Portuguese monitor corpus, CORLEX[6], ordered by alphabetical order and relative frequency, and compared. Only items with relative frequency at least two times higher in MIGRANTE. PT than in the monitor corpus were selected. The extracted lists were then manually analyzed to collect potential candidates. The frequency limit line was set to absolute corpus frequency of 70 occurrences, for verbs, 50 occurrences, for nouns, and 30 occurrences, for adjectives. Table 2 presents the first results.

Table 2. Candidates extracted from MIGRANTE.PT corpus

Feature/Part-of-Speech	Nouns	Verbs	Adjectives	TOTAL
Non-People first, specificity, tolerant/positive perspective	27	26	60	113
Non-gender neutrality	38	-	17	55
Condescension/patronizing	47	21	4	72
Inclusiveness	19	10	2	31

These first quantitative results seem to inform us that the texts have high potential to reflect discriminatory language, considering the numbers of candidates expressing bias on people's conditions and situation, gender bias and condescending/patronizing attitudes, on the one hand, versus the candidates that express inclusiveness, on the other. These results were further explored through the extraction of collocations, a feature also available in Sketch Engine. Collocations are a statistical method to identify words and/or word lemmas that co-occur with statistical relevance within a specific corpus, the keywords here being 'statistical relevance'. For the purpose of the quantitative analysis we discuss here, we extracted the top 50 collocations for several relevant candidates concerning the specificity of the domain and the first aspects of people first, specificity, tolerant/positive perspective of the nouns *migrante* (migrant), *refugiado* (refugee), *estrangeiro* (foreigner), *requerente de asilo* (asylum seeker); of the adjectives *migrante* (migrant), *refugiado* (refugee), *estrangeiro* (foreigner), and *nacional* (national); and of the verbs *assumir* (take/assume), *reconhecer* (recognize), *conceder* (grant) and *contribuir* (contribute).

A contrastive analysis of the collocations for semantically equivalent nouns and adjectives, namely the nouns and adjectives *migrante* (migrant), *refugiado* (refugee) and *estrangeiro* (foreigner) revealed that the use of the nouns refers to the non-compliance of the people first principle, although most of the collocations express integration/inclusiveness or positive ideas, as in 'migrants'/refugees' integration/contribution/health/inclusion/protection/hosting'. We also verified that there are still some expressions that convey or refer to potentially negative or discriminatory notions and with discriminative formulation such as 'removal of foreigners from national territory', 'migrants' categories', 'foreigners entry/exit'. And finally, there were also several cases of positive concepts expressed by the ideal formulation, such as 'hosting of refugee children', 'empowerment of migrant people', 'migrant/refugee people. Verb collocations express less clear results. We identified four expressions that look discriminatory – 'officially recognized', 'women assume', 'recognized as refugees', 'also contribute'; and three expressions that look inclusive – 'contribute + improvement', 'contribute + development', 'migrations contribute to'. The majority of the results need further qualitative analysis to be interpreted.

It is interesting, however, to stress that the texts under analysis come from institutions that are positively engaged in the inclusion and hosting of people on the move, which makes our results more relevant: the people aiming at enabling the inclusion of migrant and refugee people may be, unintendedly, contributing to their discrimination. One highly illustrative example is the designation *Unidade de Apoio à Vítima Migrante e de Discriminação* (Unit for Support of Migrant and Discrimination Victims)', in which people are portraited by their condition (victims) and that can be interpreted as equaling the notions of migration and discrimination (migrant and discrimination victim), as if to migrate were a similar unjust or violent situation as to be discriminated.

Following the established methodology, we extracted data on inclusiveness, namely frequencies and collocations for the nouns listed above. The results show some good practices of naming the conditions and not the individuals (e.g., *rendimentos* (income), *necessidades especiais/de proteção internacional* (special needs/need for international protection), *distúrbios mentais/perturbações* (mental disturbance), *faixas etárias* (age range) and the use of the formulae *pessoa/cidadão/indivíduo* + Adjective (people/person/individual/citizen + Adjective) and *pessoa/cidadão/indivíduo* + *com* (people/person/individual/citizen + with). However, not all cases constitute good practices of inclusive language, such as expressions stressing the contrast between national and foreign citizens and expressions employing generalizations to identify very heterogeneous groups (e.g., Asian citizens). The examples below illustrate both cases:

Table 3. Examples from suggested non-biased formulae

examples of good practices	examples of bad practices
pessoa/indivíduo migrante (migrant person/individual)	*indivíduo apátrida* (stateless person)
cidadão proveniente de (citizen coming from)	*cidadão nacional/estrangeiro* (national/foreign citizen)
cidadão natural de (citizen natural of)	*cidadão asiático* (Asian citizen)
*pessoa refugiada (*refugee/displaced person)	*cidadão/indivíduo nacional* (national citizen/individual)
pessoa menor (underage person)	
pessoa/indivíduo lésbica/gay/bissexual (lesbian/gay/bissexual individual/person)	
*cidadão com direitos (*citizen with rights)	

The collective noun with more hits was 'group', with 1914 occurrences, although not all collocations were relevant for inclusiveness (e.g., 'in the group', 'by group'), followed by 'community', 775 occurrences. 'Class' occurred only 38 times, most of them referring to income distribution ('middle class'). For these types of expressions, collocations show several good practice results such as 'migrant/roma/religious/educational/ethnic community/groups'; 'political/medical class'; 'vulnerable/professional/age group'.

The results on verbal proxemics, i.e., 1st person use, covering personal pronouns, possessives, and verbs in 1st person, both singular and plural (Carreira, 1997; Carreira, 2008; Íñigo-Mora, 2004), indicate that this is not a salient strategy for these texts. Table 4 shows the contrastive distribution of these forms in MIGRANTE.PT corpus and in the CORLEX corpus, the monitor corpus.

Table 4. 1st person forms distribution

	MIGRANTE.PT	CORLEX
determinants & pronouns	649 tokens 436,6 per million	57 506 tokens 7 279,2 per million
verb forms	5 494 tokens 3 924,3 per million	135 673 tokens 17 173,8 per million

Extracting more directed data, in particular the use of 1st person pronouns to establish the dynamics of 'us vs. them', shows that the co-occurrence of 'us/we' + 'them' in restricted contexts, i.e., in which the pronouns must co-occur within a window of 0-10 words, in MIGRANTE.PT is not frequent. We extracted 57 occurrences and, of the relevant contexts retrieved two refer to this opposition as an example of bad practices and xenophobic trends, and two refer to the direct narration of personal experiences, the 'us' being the people on the move. The examples below illustrate these cases. In 1., the 'we' (*nós*) refer to people on the move, in Portugal and fully integrated in the labour market and tax system, while 'they' (*eles*) refers to people living in Portugal, assumed with Portuguese nationality, and it reports a xenophobic attitude. In example 2., 'we' (*nós, erámos*) refers to people from Portugal and 'they' (*eles*) refers to people coming from outside Europe. The sentence expresses a clear separation from xenophobic feelings and a direct empathic relation with people on the move, coming from personal experience.

1. *temos de fazer os descontos e eles entendiam que nós "roubávamos-lhe" o que lhes pertencia* (we have to make the discounts and they understood that we "stole" what belonged to them from them
2. *Eles não são europeus, têm uma fisionomia diferente mas nós também emigrámos e éramos diferentes dos povos do centro da Europa* (They are not European, they have a different physiognomy, but we also emigrated and we were different from the people of central Europe).

Discussing the Results

The analysis carried out enables us to conclude that the extraction of quantitative data from language corpora, based on an informed dataset, can trigger important cues on how to assess inclusiveness, in particular:

- specific vocabulary analysis through keyword and specialized lexicon extraction.
- context determination through concordances and, most importantly, through collocations extraction and analysis.
- retrieval of specific language-dependent phenomena using linguistic features annotated in language corpora (namely, fine-grained Part-of-Speech and morphological annotation and lemmatization).

However, what is interesting to notice is that these measures can go either way: either they provide information on inclusive texts, or they show evidence of non-inclusive language. When systematizing results to find patterns and general trends, one finds contradictory information. For instance, the presence of specific lexical structures such as neutral expressions for person (e.g., *pessoa/indivíduo/cidadão* (person/individual/citizen)) + Adjective are a cue for inclusiveness, but this pattern often expresses non-inclusive notions, depending on the co-occurrent adjective (e.g., 'foreign citizen'). Also, general trends seem to impact the analysed texts, even though these are products of public and private institutions with special responsibilities and goals regarding successful inclusion/integration of migrant people. The two noticeable cases concern gender neutrality and condescension/patronizing aspects: gender bias is also reflected in these texts (e.g. *mulheres grávidas* (pregnant women), *trabalhadores* (morphologically masculine workers), *os pais* (morphologically masculine parents), and a general trend for stressing the hard conditions of some people, such as the proper name of the *Unidade de Apoio à Vítima Migrante e de Discriminação* (Unit for Support of Migrant and Discrimination Victims) is such an exemplar case.

The results show us that more must be done in terms of an automatic assessment of inclusiveness in terms of gathering more fine-grained features to inform the analysis. The use of balanced and processed language corpora is an advantage and can easily show us examples of inclusive/non inclusive language but the basis on which this method can be productive has to be improved. The texts must be thoroughly and naively scrutinized to identify other specific features related to inclusiveness, going much further than the general indications inclusive guidelines can inform.

ASSESSING AND SELECTING MATERIALS FOR INCLUSIVE HOST LANGUAGE TEACHING

The analysis, informed by current inclusiveness guidelines, of Portuguese official texts targeting migrant and refugee people presented so far leaves several overlooked issues still to be tackled. Besides the obvious language-dependent features that necessarily must be taken into account, there are other linguistic values not so straightforwardly assessable that require further qualitative analysis. This will lead to a better understanding of what is at stake in the assessment and selection of materials for inclusive host language teaching, while showing future work directions.

Accounting for Semantic Values Relevant for Inclusiveness Assessment

In addition to the quantitative analysis carried out above, it is also important to perform a qualitative analysis to highlight some of the semantic values that the expressions acquire in context of use. To this end, we have chosen the three most commonly used names for identifying people – ‘migrant’, ‘refugee’ and ‘foreigner’ – and observed the following phenomena.

Although refugees and immigrants may be used equivalently, especially in the media, they correspond to different terms, as defined by UNHCR (United Nations High Commissioner for Refugees, n.d.):

Refugees *are persons fleeing armed conflict or persecution. Refugees are defined and protected in international law. The 1951 Refugee Convention and its 1967 Protocol as well as other legal texts, such as the 1969 OAU Refugee Convention, remain the cornerstone of modern refugee protection. The legal principles they enshrine have permeated into countless other international, regional, and national laws and practices. The 1951 Convention defines who is a refugee and outlines the basic rights which States should afford to refugees.*

Migrants *choose to move not because of a direct threat of persecution or death, but mainly to improve their lives by finding work, or in some cases for education, family reunion, or other reasons. Unlike* *refugees who cannot safely return home, migrants face no such impediment to return. If they choose to return home, they will continue to receive the protection of their government.*

Confusing them leads to problems for both populations. In fact, countries deal with migrants under their own immigration laws and deal with refugees through norms of refugee protection, defined in both national legislation and international law. In texts, we find confusion and fluctuation in the choice of terms:

1. *At the level of the Lisbon Municipality's social responses for particularly vulnerable groups of migrants is the intervention in the area of asylum seekers and persons with international protection status, irregular foreigners and homeless immigrant people.*
2. *CML created, in 2015, the Municipal Lisbon Refugee Reception Programme (PMAR Lx). This Programme arises in response to the commitment made by the Municipality of Lisbon to welcome around 10% of the national total of the contingent that the European Commission assigned to Portugal under its relocation policies, i.e., around 500 refugee persons.* (translated from CM_ACM_LisboaPlanoMunicipalIntegracao_Lisboa2018-2020.txt)

Example 1. is taken from the text *Plano Municipal para Integração dos Migrantes em Lisboa* and an assimilation between migrant and refugee is observed. In fact, the text refers to refugees and defines them as a particularly vulnerable group of migrants. With the choice of the adjective vulnerable, a condescension value is created, perpetuating a bias. However, in example 2., we observe the lexical choice *refugees* (refugiados) and *refugee people* (pessoas refugiadas) more appropriate, which proves to be more inclusive. It should be noted that this phenomenon occurs occasionally in the text and not in its entirety. Still on the amalgamation of refugees and migrants, note these examples extracted from news:

3. *Geneva, 7 September 2015 (UNHCR/CPR) – At least 124,000 of the 225,000 refugees and other migrants who arrived in Europe from January to July 2015 via the Mediterranean Sea landed in Greece, mainly on the islands of Lesbos, Kos, Chios, Samos and Lero.*
4. *Zagreb, 27 October 2015 (UNHCR) On Sunday, a convoy from Tovarnik, a Croatian town on the border with Serbia, carrying 1,800 refugees and other migrants arrived in Cakovec, on the border with southern Slovenia but passage was barred by the police (…).*
5. *Lisbon, December 11 (CPR) – Melissa Fleming, UNHCR spokesperson, was interviewed by the BBC this Thursday, Dec. 10, about the refugees and other migrants, mainly Syrians, who continue to make the Aegean Sea crossing towards Europe.*

(Excerpts translated from news texts CPR_Notícias_siteCPR_2015-2020, MIGRANTE.PT)

6. *Compared to the population with Portuguese nationality, in 2011, the foreign population was more dependent on labour income (+18.5 percentage points) and had a dependence on pensions/pensions of less than 5%.* (Excerpt translated from the institutional text CM_ACM_LisboaPlanoMunicipalIntegracao_Lisboa2018-2020.txt, MIGRANTE.PT)

The confusion between migrant and refugee people is quantitatively more common in dissemination texts (news and media) than in legal texts, as shown in the previous examples. Also, another phenomenon that does not contribute to the creation of inclusiveness and the aggregation of people that was detected concerns the dissociation between people of Portuguese nationality and other nationalities, illustrated in example 3. This example reflects the creation of two distinct poles and a high verbal proxemic – Portuguese population versus foreign population – without establishing an interrelationship between them. Although stated in a municipality plan for integration, this dissociation does not guarantee or promote integration or inclusion.

Selecting Appropriate Teaching Material

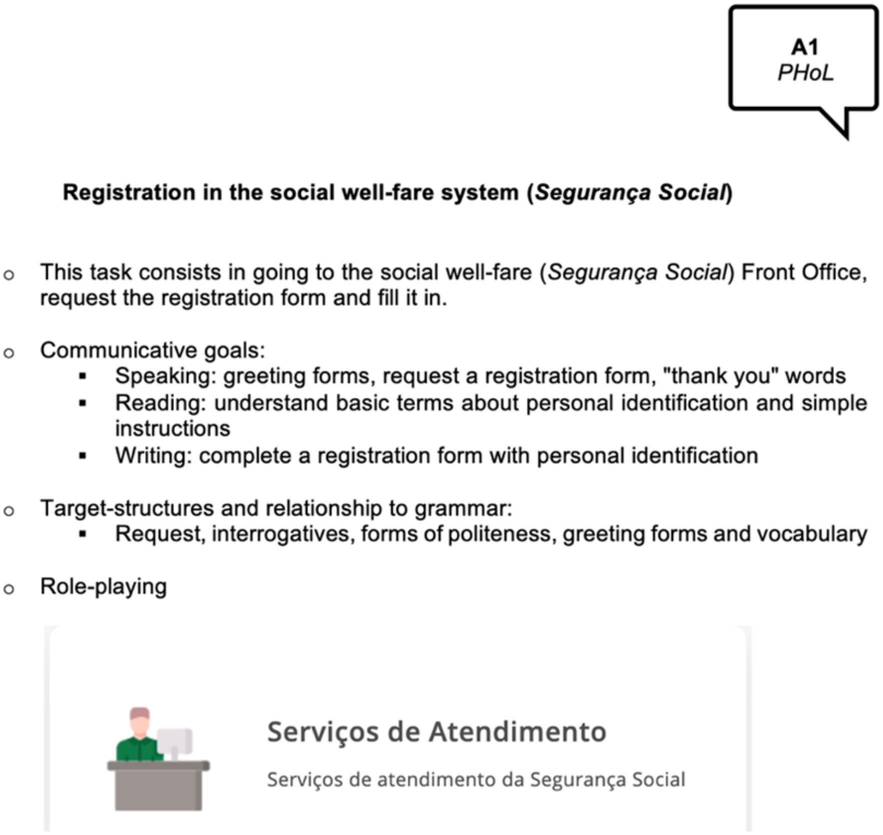
The selection of appropriate teaching materials allowing for, or promoting, inclusion and aggregation requires more than a simple wordlist to be used. The issues discussed thus far inform a set of recommendations that account for the principles of people first; specificity – not only in terms of the people's characterization, but also regarding their specific conditions and needs; tolerant/positive perspective – including gender, ethnic and cultural sensibility; neutrality in the use of gender, as well as ethnic, religious or other individual or sociocultural conditions; openness to inclusion values by promoting the active insertion of themes, needs, topics of concern, traditions, etc., in the teaching/learning activities and situations, going beyond strict neutrality. These are formulated in the following 10 recommendations for an inclusive host language teaching:

1. Characterize the target audience in terms of age, gender, language, SES, and sociocultural background before class. If a needs analysis is not possible before the first class, do that in the first class. This should be done, preferably, in ice-breaking activities where learners are approached with warmth. Use a mediator or a mediation language, if necessary. Gather as much information as possible.
2. Design inclusive tasks and activities using inclusive and diverse themes and vocabulary (multi-ethnic, multi-gendered and multicultural; refer to "people with…", "people that are…", ...).
3. Avoid sensitive themes and vocabulary (e.g., bullfights when teaching traditions, sacred animals when teaching food).
4. Avoid using audiovisual material that contains sensitive images (e.g., videos that are ethnically biased in the people they show).
5. Adapt texts for proficiency level, but also for vocabulary and grammatical constructions that refer to gender, ethnicity, religion, disabilities.
6. Illustrate language teaching activities with pictures and images that are inclusive and culturally varied (multi-ethnic, multi-gendered and multicultural – include images of people of different colors, religious outfits, ages, genders).
7. When approaching religiously marked themes and vocabulary (e.g., Christmas or Easter), ask the learners whether there is a similar celebration in their culture; be sensitive to sensitive themes (family, money, ethnicity, religion, disabilities).
8. Avoid expressions such as "we/us/our culture/religion/gender/race vs. they/you/your culture/religion/gender/race"). Use instead "Portuguese culture", "Catholicism", "people (meaning women and men)", "people that are/came from…" (meaning ethnicities).
9. Do not take a cultural, educational, religious, or socioeconomic background for granted (not everyone knows everything). Adapt a cultural, educational, or religious background taking in account the students' different background and the communicative needs. The use of cultural stereotypes in teaching materials can transform the classroom into a general culture contest that can segregate students from different backgrounds.

10. Do not be neutral, just do not be biased.

In the following pages, an example of a classroom activity with a task for Portuguese as a Host Language targeting A1 learners is shown (Figures 2 to 4). In this task, the use of several expressions that suggest exclusion and bias is observed. First, “cidadão estrangeiro” (foreign citizen), an expression that appears multiple times, implies that the learner is a *citizen*. The same happens for “Cartão do Cidadão estrangeiro” (foreign ID card). In an inclusive host language teaching situation, the material could undergo adaptations, such as “estrangeiro” being replaced by “from other countries” and “cidadãos espanhóis” or “cidadão português” (Spanish citizens or Portuguese citizen) could be replaced by “people from Spain” and “people from Portugal”, respectively. In addition, “refugiado” (refugee) is used disjunctively with “beneficiário de proteção subsidiária” (beneficiary of subsidiary protection), which is more inclusive than simply using *refugee*, but still the use of expressions such as “refugees”, “people applying for international protection”, “beneficiary of international protection” seems preferable and less condescending. “Para todas as pessoas” (to all the people) should replace “todos” (all, ‘generic masc. pl.’), preventing the masculine bias. The use of masculine plural nouns to refer to a collective of men and a collective of men and women is, in fact, a property of the language that is undergoing change due to language inclusion policies. More and more, institutional guidelines recommend using “todos e todas” (all ‘masc. pl.’ and all ‘fem. pl.’) or “todas as pessoas” (all the people) to refer to collectives of men and women. Finally, vocabulary such as “Filiação” (filiation/names of parents) should be defined and explained.

Figure 2. TBLT activity script



A1
PHoL

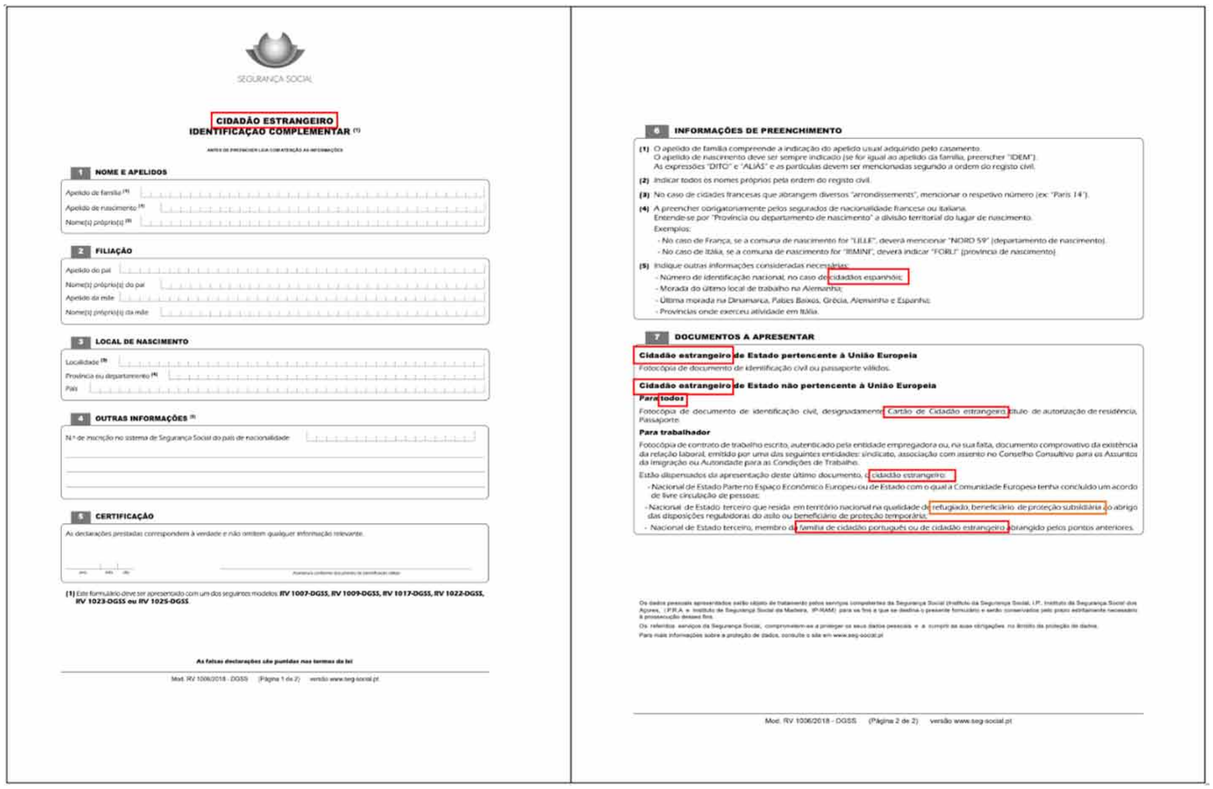
Registration in the social well-fare system (*Segurança Social*)

- This task consists in going to the social well-fare (*Segurança Social*) Front Office, request the registration form and fill it in.
- Communicative goals:
    - Speaking: greeting forms, request a registration form, "thank you" words
    - Reading: understand basic terms about personal identification and simple instructions
    - Writing: complete a registration form with personal identification
- Target-structures and relationship to grammar:
    - Request, interrogatives, forms of politeness, greeting forms and vocabulary
- Role-playing

Serviços de Atendimento

Serviços de atendimento da Segurança Social

Figure 3. Real form extracted from Seguranca Social, n.d.



SEGURANÇA SOCIAL

CIDADÃO ESTRANGEIRO
IDENTIFICAÇAO COMPLEMENTAR (1)

ANTES DE PREENCHER LEIA COM ATENÇÃO AS INFORMAÇÕES

1 NOME E APELIDOS

Apelido de família (1)
Apelido de nascimento (1)
Nome(s) próprio(s) (2)

2 FILIAÇÃO

Apelido do pai
Nome(s) próprio(s) do pai
Apelido da mãe
Nome(s) próprio(s) da mãe

3 LOCAL DE NASCIMENTO

Localidade (3)
Província ou departamento (4)
País

4 OUTRAS INFORMAÇÕES (5)

N.º de inscrição no sistema de Segurança Social do país de nacionalidade

5 CERTIFICAÇÃO

As declarações prestadas correspondem à verdade e não omitem qualquer informação relevante.

ano mês dia

Assinatura conforme documento de identificação válido

(1) Este formulário deve ser apresentado com um dos seguintes modelos: **RV 1007-DGSS, RV 1009-DGSS, RV 1017-DGSS, RV 1022-DGSS, RV 1023-DGSS ou RV 1025-DGSS.**

As falsas declarações são punidas nos termos da lei

Mod. RV 1006/2018 - DGSS (Página 1 de 2) versão www.seg-social.pt

6 INFORMAÇÕES DE PREENCHIMENTO

(1) O apelido de família compreende a indicação do apelido usual adquirido pelo casamento.
O apelido de nascimento deve ser sempre indicado (se for igual ao apelido da família, preencher "IDEM").
As expressões "DITO" e "ALIÁS" e as partículas devem ser mencionadas segundo a ordem do registo civil.

(2) Indicar todos os nomes próprios pela ordem do registo civil.

(3) No caso de cidades francesas que abrangem diversos "arrondissements", mencionar o respetivo número (ex: "Paris 14").

(4) A preencher obrigatoriamente pelos segurados de nacionalidade francesa ou italiana.
Entende-se por "Província ou departamento de nascimento" a divisão territorial do lugar de nascimento.
Exemplos:
- No caso de França, se a comuna de nascimento for "LILLE", deverá mencionar "NORD 59" (departamento de nascimento).
- No caso de Itália, se a comuna de nascimento for "RIMINI", deverá indicar "FORLI" (província de nascimento).

(5) Indique outras informações consideradas necessárias:
- Número de identificação nacional, no caso de cidadãos espanhóis;
- Morada do último local de trabalho na Alemanha;
- Última morada na Dinamarca, Países Baixos, Grécia, Alemanha e Espanha;
- Províncias onde exerceu atividade em Itália.

7 DOCUMENTOS A APRESENTAR

Cidadão estrangeiro de Estado pertencente à União Europeia

Fotocópia de documento de identificação civil ou passaporte válidos.

Cidadão estrangeiro de Estado não pertencente à União Europeia

Para todos

Fotocópia de documento de identificação civil, designadamente Cartão de Cidadão estrangeiro, título de autorização de residência, Passaporte.

Para trabalhador

Fotocópia de contrato de trabalho escrito, autenticado pela entidade empregadora ou, na sua falta, documento comprovativo da existência da relação laboral, emitido por uma das seguintes entidades: sindicato, associação com assento no Conselho Consultivo para os Assuntos da Imigração ou Autoridade para as Condições de Trabalho.

Estão dispensados da apresentação deste último documento, o cidadão estrangeiro:
- Nacional de Estado Parte no Espaço Económico Europeu ou de Estado com o qual a Comunidade Europeia tenha concluído um acordo de livre circulação de pessoas;
- Nacional de Estado terceiro que resida em território nacional na qualidade de refugiado, beneficiário de proteção subsidiária ao abrigo das disposições reguladoras do asilo ou beneficiário de proteção temporária;
- Nacional de Estado terceiro, membro da família de cidadão português ou de cidadão estrangeiro abrangido pelos pontos anteriores.

Os dados pessoais apresentados estão sujeitos de tratamento pelos serviços competentes da Segurança Social (Instituto da Segurança Social, I.P., Instituto da Segurança Social dos Açores, I.P.R.A. e Instituto de Segurança Social da Madeira, IP-RAM) para os fins a que se destina o presente formulário e serão conservados pelo prazo estritamente necessário à prossecução desses fins.

Os referidos serviços da Segurança Social, comprometem-se a proteger os seus dados pessoais e a cumprir as suas obrigações no âmbito da proteção de dados.

Para mais informações sobre a proteção de dados, consulte o site em www.seg-social.pt

Mod. RV 1006/2018 - DGSS (Página 2 de 2) versão www.seg-social.pt

Figure 4. TBLT activity assessment

- Assess your performance:

Was this task completed successfully?		*Did you achieve the main communicative goal?*		*Strengths*	*Weaknesses*
Yes □	No □	Yes □	No □	• ____ • ____ • ____ • ____ • ____ • ____	• ____ • ____ • ____ • ____ • ____ • ____

- Identify problems and clarify doubts
- Repeat the as many times as needed
- Practice and extend to similar contexts - e.g., request registration in the tax system (*Autoridade Tributária*); request a health care facility (*Centro de Saúde*)

FINAL REMARKS AND FUTURE RESEARCH DIRECTIONS: FEATURING INCLUSIVENESS AND AGGREGATION IN TEXTS

Language knowledge and communication skills are a key factor in promoting social and labour market integration, in giving migrant people access to rights and services, including them in the host society, and empowering them to become citizens. In order to contribute to the improvement of reception conditions for migrant people, through language, this work approached two objectives:

1. Starting from concrete texts circulating in society, with different communicative characteristics and purposes (informative, legal, dissemination and technical), addressed to migrant and refugee people, we analysed and detected linguistic features of inclusivity and bias.
2. Building on the analysis work, we provided insights and recommendations on how to select and adapt these real texts for use in task-based language teaching for inclusive host language teaching.

According to what has been demonstrated throughout this paper, the model and the methodology used – qualitative/quantitative analysis through the use of NLP tools – showed great relevance and potential for the detection of biases and degrees of inclusion by language. It was clear that inclusive and non-inclusive features were detected in the corpus, which allowed for drawing up a list of recommendations for the selection of suitable materials for language teaching and contribute for inclusive language teaching.

This work shows how quantitative analysis of the corpus provides important data regarding inclusiveness. However, as we mentioned earlier, this analysis can have different meanings depending on the context in which it takes place. It seems important to us to add new layers of analysis, such as the activity in which the text is produced or its communicative purpose, which can configure a usage pattern. This way, the annotation of semantic elements, although problematic in methodological terms, can add another angle of analysis. One of the semantic elements that can be added is the perception on neutrality, inclusiveness, and aggregation that readers have of the collocations extracted from the corpus, and of the texts themselves, by means of a specific annotation task to build training corpus for machine learning purposes. The key idea is to gather robust intuitive information, through a controlled consultation of the target audience – migrant and refugee people – specifically designed to cover all stakeholders in diverse situations and profit from the potentially neutral approach that machine-learning systems can provide. As noted earlier, however, special attention must be given to the input material gathered so that bias is not inadvertently passed on to the automated system.

The work depicted in this chapter, as well as the future lines of research put forth, will add knowledge about language use for teaching purposes not covered by current NLP analysis, sustaining future work on inclusiveness/aggregation annotation for text mining/machine learning in a more efficient and productive way.

ACKNOWLEDGMENT

This research was supported by the Portuguese national funding through the FCT – Portuguese Foundation for Science and Technology as part of the project of CLUNL, School of Social Sciences and Humanities, NOVA University Lisbon, 1069-061 Lisbon, Portugal (UIDB/LIN/03213/2020 and UIDP/

LIN/03213/2020), and by FCT – KRUse PhD grant PD/BD/142789/2018 and grant FCSH/CLUNL/UIDB/03213/2020 /1BI.

REFERENCES

Abranches, G. (2009). *Guia para uma Linguagem Promotora da Igualdade entre Mulheres e Homens na Administração Pública*. Comissão para a Cidadania e Igualdade de Género Presidência do Conselho de Ministros. https://www.cig.gov.pt/wp-content/uploads/2015/11/Guia_ling_mulhe_homens_Admin_Publica.pdf

ACM. (n.d.). Retrieved from https://www.acm.gov.pt/pt/-/politicas-locais-para-acolhimento-e-integracao-dos-imigrantes

Adam, J. M. (2005). *La Linguistique Textuelle. Introduction à l'analyse textuelle des discours*. Armand Colin.

Avineri, N., Graham, L. R., Johnson, E. J., Robin, C. R., & Jonathan, R. (Eds.). (2019). *Language and Social Justice in Practice*. Routledge.

Beukeboom, C. J., & Burgers, C. (2017). Linguistic bias. In H. Giles & J. Harwood (Eds.), *Oxford Encyclopedia of Intergroup Communication*. Oxford University Press. doi:10.1093/acrefore/9780190228613.013.439

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*(4), 243–257. doi:10.1093/llc/8.4.243

Bleakley, H., & Chin, A. (2010). Age at arrival, English proficiency, and social assimilation among US immigrants. *American Economic Journal. Applied Economics*, *2*(1), 165–192. doi:10.1257/app.2.1.165 PMID:20119509

Blodgett, S. L. (2021). *Sociolinguistically Driven Approaches for Just Natural Language Processing* (PhD Dissertation). University of Massachusetts Amherst.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, *29*, 4349–4357.

Borchmann, S., Levisen, C., & Schneider, B. (2019). Linguistics as a biased discipline: Identifications and interventions. *Language Sciences*, *76*, 101218. Advance online publication. doi:10.1016/j.langsci.2018.12.003

Bronckart, J.-P. (1997). *Activité langagière, textes et discours. Pour un interactionnisme discursif*. Delachaux et Niestlé.

Caels, F. (2016). *Guia para o ensino do Português enquanto Língua de Acolhimento (PLA) no contexto da Educação Não Formal (ENF)*. ACM. https://www.acm.gov.pt/documents/10181/222893/Guia_ENF_vf_portal.pdf/524f6621-ace4-4da2-b769-7bf0c6341e2a

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. doi:10.1126cience.aal4230 PMID:28408601

Carliner, G. (1981). Wage differences by language group and the market for language skills in Canada. *The Journal of Human Resources*, *16*(3), 384–399. doi:10.2307/145627

Carreira, M. H. (1997). *Modalisation linguistique en situation d'interlocution: proxémique verbale et modalités en portugais*. Peter- Lang.

Carreira, M. H. (2008). Deixis e proxémica verbal: percursos enunciativos e processos discursivos. In *O Fascínio da linguagem. Actas do Colóquio de homenagem a Fernanda Irene Fonseca* (pp. 297–308). CLUP/FLUP. https://ler.letras.up.pt/uploads/ficheiros/6711.pdf

CDCC – Council for Cultural Cooperation. (2000). Project "Education for democratic citizenship". Basic concepts and core competences for education for democratic citizenship. University of Geneva.

CES – Conselho Económico e Social. (2021). *Manual de Linguagem Inclusiva*. Conselho Económico e Social. https://www.ces.pt/storage/app/uploads/public/60a/bcf/01a/60abcf01a49a6966725992.pdf

Chen, M. K. (2013). The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *The American Economic Review*, *103*(2), 690–731. doi:10.1257/aer.103.2.690 PMID:29524925

Chiswick, B. R., & Miller, P. W. (2014). *International Migration and the Economics of Language*. IZA DP No. 7880. http://repec.iza.org/dp7880.pdf

CoE – Council of Europe. (2018). *Employment and Social Developments in Europe 2018*. Publications Office of the European Union.

COSWL & LSA – Committee on the Status of Women in Linguistics & Linguistic Society of America. (2016). *Guidelines for Inclusive Language*. https://www.linguisticsociety.org/resource/guidelines-inclusive-language

Cotos, E. (2017). Language for specific purposes and corpus-based pedagogy. In C. A. Chapelle & S. Sauro (Eds.), *The handbook of technology and second language teaching and learning* (pp. 248–264). John Wiley & Sons, Inc. doi:10.1002/9781118914069.ch17

Coulmas, F. (2017). Language and Society: Historical Overview and the Emergence of a Field of Study. In O. García, N. Flores, & M. Spotti (Eds.), *The Oxford Handbook of Language and Society*. Oxford University Press.

De Houwer, J. (2006). What Are Implicit Measures and Why Are We Using Them? In R. W. Wiers & A. W. Stacy (Eds.), *Handbook of implicit cognition and addiction* (pp. 11–28). Sage Publications, Inc., doi:10.4135/9781412976237.n2

Del Valle, J. (Ed.). (2007). *La lengua,¿ patria común? Ideas e ideologías del español* (Vol. 17). Iberoamericana Editorial.

EC – European Commission. (2020). *Action plan on Integration and Inclusion 2021-2027*. https://ec.europa.eu/home-affairs/sites/default/files/pdf/action_plan_on_integration_and_inclusion_2021-2027.pdf

ECRI – European Commission against Racism and Intolerance. (n.d.). *Integration and Inclusion*. https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/integration-and-inclusion

EIGE – European Institute for Gender Neutrality. (2019). *Toolkit on Gender-sensitive Communication A resource for policymakers, legislators, media and anyone else with an interest in making their communication more inclusive*. Publications Office of the European Union. https://eige.europa.eu/sites/default/files/20193925_mh0119609enn_pdf.pdf

Ellis, R. (2017). Position paper: Moving task-based language teaching forward. *Language Teaching*, *50*(4), 507–526. doi:10.1017/S0261444817000179

Elsod, A. & Marques, M. (Coord.). (2019). *Ask the People: a consultation of migrants and refugees*. EMAB, Open Society Foundations.

EU – European Commission. (2016). *English Style Guide A handbook for authors and translators in the European Commission.* https://ec.europa.eu/info/sites/default/files/styleguide_english_dgt_en.pdf

EU – European Union. (2020). *Action plan on Integration and Inclusion 2021-2027*. https://ec.europa.eu/home-affairs/sites/default/files/pdf/action_plan_on_integration_and_inclusion_2021-2027.pdf

EU – European Union. (n.d.)"Gender-neutral language. In *Interinstitutional Style Guide*. Publications Office of the European Union. http://publications.europa.eu/code/en/en-4100600en.htm

Fairclough, N. (1989/1994). *Language and Power*. Longman.

Fairclough, N. (2003). *Analysing Discourse: Textual Analysis for Social Research*. Routledge. doi:10.4324/9780203697078

Foucault, M. (1970). *The Order of Things: an Archaeology of the Human Sciences*. Tavistock.

Foucault, M. (1972). *The Archaeology of Knowledge*. Tavistock.

GBC – Government of British Columbia. (n.d.). *Guidelines on using inclusive language in the workplace*. https://www2.gov.bc.ca/assets/gov/careers/all-employees/working-with-others/words-matter.pdf

GMSA. (2020). *Inclusive Language Guide*. https://www.gsma.com/aboutus/wp-content/uploads/2020/11/GSMA-Inclusive-Language-Guide_2020.pdf

Guven, C., & Islam, A. (2015). Age at migration, language proficiency, and socioeconomic outcomes: Evidence from Australia. *Demography*, *52*(2), 513–542. doi:10.100713524-015-0373-6 PMID:25749486

Íñigo-Mora, I. (2004). On the use of the personal pronoun we in communities. *Journal of Language and Politics*, *3*(1), 27–52. doi:10.1075/jlp.3.1.05ini

Jaeger, F. N., Pellaud, N., Laville, B., & Klauser, P. (2019). The migration-related language barrier and professional interpreter use in primary health care in Switzerland. *BMC Health Services Research*, *19*(1), 1–10. doi:10.118612913-019-4164-4 PMID:31248420

Long, M. (2014). *Second language acquisition and task-based language teaching*. Wiley-Blackwell.

Mann, T. C., Kurdi, B., & Banaji, M. R. (2020). How effectively can implicit evaluations be updated? Using evaluative statements after aversive repeated evaluative pairings. *Journal of Experimental Psychology. General*, *149*(6), 1169–1192. doi:10.1037/xge0000701 PMID:31670568

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice*. Cambridge University Press.

McManus, W., Gould, W., & Welch, F. (1983). Earnings of Hispanic men: The role of English language proficiency. *Journal of Labor Economics*, *1*(2), 101–130. doi:10.1086/298006

Miranda, A., & Zhu, Y. (2013). English deficiency and the native–immigrant wage gap. *Economics Letters*, *118*(1), 38–41. doi:10.1016/j.econlet.2012.09.007

OECD – Organization for Economic Co-operation and Development. (2018). *International Migration Outlook 2018* (42nd ed.). OECD Publishing.

OECD – Organization for Economic Co-operation and Development. (2021). Use inclusive language to refer to youth with migrant parents. In *Young People with Migrant Parents*. OECD Publishing., doi:10.1787/57ff5c96-

OSHU – Oregon Health &Science University. (2021). *Inclusive Language Guide. An evolving tool to help OHSU members learn about and use inclusive language*. OSHU. https://www.ohsu.edu/sites/default/files/2021-03/OHSU%20Inclusive%20Language%20Guide_031521.pdf

Pastor, G. C., & Seghiri, M. (2009). Virtual corpora as documentation resources: Translating travel insurance documents. *Corpus use and translating: Corpus use for learning to translate and learning corpus use to translate, 82*, 75-107. doi:10.1075/btl.82.07cor

Portes, A., & Rumbaut, R. G. (2001). *Legacies: The Story of the Immigrant Second Generation*. University of California Press and Russell Sage Foundation.

Rastier, F. (2001). *Arts et sciences du texte*. PUF. doi:10.3917/puf.rast.2001.01

Robinson, P. (2011). Task-based language learning: A review of issues. *Language Learning*, *61*, 1–36. doi:10.1111/j.1467-9922.2011.00641.x

Rosa, J. (2019). Contesting Representations of Migrant “Illegality” through the Drop the I-Word Campaign: Rethinking Language Change and Social Change. In N. Avineri, L. R. Graham, E. J. Johnson, R. C. Riner, & J. Rosa (Eds.), *Language and Social Justice in Practice* (pp. 35–43). Routledge. doi:10.4324/9780429199240

Seguranca Social. (n.d.). *Cidadão Estrangeiro Identificação Complementar*. Retrieved from https://www.seg-social.pt/documents/10152/38806/RV_1006_DGSS/d40ab4c2-9080-4bf9-a8ae-a772b43edc2b

Serrão, C., Martins, T., & Rocha, R. M. (2020). *Guia para uma comunicação inclusiva.* Instituto Politécnico do Porto. https://www.ipp.pt/comunidade/responsabilidade_social/comunicacao_inclusiva/copy2_of_GUIAINCLUSAOWEB.pdf

Tortajada, I., Comas d'Argemir, D., Muixí, M., Martínez, R., & Guarro, B. (2013). *Guia de llenguatge inclusiu Immigració, racisme i xenofòbia.* https://www.mesadiversitat.cat/sites/default/files/2017-11/guia_breu_llenguatge_inclusiu.pdf

UN – United Nation Office. (2019). https://www.ungeneva.org/sites/default/files/2021-01/Disability-Inclusive-Language-Guidelines.pdf

UN Women – United Nation Women. (n.d.). *Gender-inclusive language guidelines (English). Promoting gender equality through the use of language.* https://www.unwomen.org/-/media/headquarters/attachments/sections/library/gender-inclusive%20language/guidelines-on-gender-inclusive-language-en.pdf?la=en&vs=2129

UNESCO – United Nations Educational, Scientific and Cultural Organization. (1994). *The Salamanca Statement and Framework for Action on Special Needs Education.* https://unesdoc.unesco.org/ark:/48223/pf0000098427

United Nations High Commissioner for Refugees. (n.d.). *UNHCR viewpoint: 'refugee' or 'migrant' – which is right?* UNHCR. Retrieved from https://www.unhcr.org/news/latest/2016/7/55df0e556/unhcr-viewpoint-refugee-migrant-right.html

Van Dijk, T. A. (2014). Discourse, cognition, society. *The discourse studies reader: Main currents in theory and analysis, 388,* 121-146. http://www.discourses.org/OldArticles/Ideology.pdf

Voloshinov, V. (1929). *Marxismo e Filosofia da linguagem*. Hucitel.

ADDITIONAL READING

Avineri, N., Graham, L. R., Johnson, E. J., Robin, C. R., & Jonathan, R. (Eds.). (2019). *Language and Social Justice in Practice*. Routledge.

Beukeboom, C. J., & Burgers, C. (2017). Linguistic bias. In H. Giles & J. Harwood (Eds.), *Oxford Encyclopedia of Intergroup Communication*. Oxford University Press., doi:10.1093/acrefore/9780190228613.013.439

Cotos, E. (2017). Language for specific purposes and corpus-based pedagogy. In C. A. Chapelle & S. Sauro (Eds.), *The handbook of technology and second language teaching and learning* (pp. 248–264). John Wiley & Sons, Inc. doi:10.1002/9781118914069.ch17

Ellis, R. (2017). Position paper: Moving task-based language teaching forward. *Language Teaching*, *50*(4), 507–526. doi:10.1017/S0261444817000179

KEY TERMS AND DEFINITIONS

Aggregation: The act or process of bringing together individuals in a collective organization or a whole.

Corpus Linguistics: Subfield of linguistics that studies language from large sets of electronically stored language data, collected and compiled from real world usage using specific criteria to assure representativeness, and analysed using dedicated NLP tools.

Host Language: The official and/or communication language spoken in a given geopolitical territory.

Inclusiveness: The property of actions, situations, or texts (written or spoken) that ensures that the needs, specific challenges and conditions of all the people involved are accounted for.

Linguistic Bias: The set of linguistic expressions – words, phrases, or constructions – that denote social, religious or other type of bias, either intentionally or by cultural and historical reasons.

Machine Learning: Subfield of computer science that studies and develops computation algorithms that improve automatically through experience, based on sample data.

Task-Based Language Teaching: Evidence-based approach to language teaching, including curriculum design and the development of teaching materials, using actual and real communicative needs and focusing on the achievement of a specific goal.

ENDNOTES

[1] https://www.coe.int/en/web/language-policy/overview

[2] https://bootcat.dipintra.it/

[3] https://www.sketchengine.eu/

[4] https://www.sketchengine.eu/portuguese-freeling-part-of-speech-tagset/

[5] Fifteen guidelines and studies were analysed: 11 international (OECD, 2021; OSHU, 2021; COSWL & LSA, 2016: UN, 2019; UN Women, n.d.; EIGE, 2019; GMSA, 2019; EU, ongoing; EC, 2016; GBC, n.d.; Torjada et al., 2013); 4 national (CES, 2021; Serrão et al., 2020; Caels, 2016; Abranches, 2009).

[6] CORLEX is a balanced sub corpus from *CRPC – Reference Corpus of Contemporary Portuguese*, with 7,912,295 tokens. For CORLEX corpus description, see http://clul.ulisboa.pt/en/recurso/multifunctional-computational-lexicon-contemporary-portuguese.