***I. Chiari, M. Bader, A. Salem and L. Squillante***

**Using Corpora in Building a Multilingual**

**Glossary of Migration**

**Abstract.** This paper describes the different steps in which corpora have been involved in the construction of a multilingual glossary of migration terminology that is meant to represent different stratification of the lexicon pertaining the migration domain starting from a set of corpora of parallel and comparable corpora in Italian English and Arabic. The Language on the Fly project is a monodirectional multilingual platform aimed at representing a prototype resource for glossaries that include a stratification of terminology varied in scope: international, national, and local.

**Keywords.** corpus-based lexicography, multilingual glossary, Italian, English, Arabic.

# 1. Introduction

***1.1. Terminology of Migration and its purposes and addressee***

Building an updated and reliable glossary of migration is a challenge and at the same time a duty at a time when mass movements, especially towards Europe, oblige us to face daily interactions with a vast number of extensively diverse languages, whose origin countries have customs, cultures, legal and administrative systems sometimes significantly different from those shared in the European framework. Thus, a glossary of migration has both a linguistic and a civil purpose: making accessible, understandable, and comparable terminology that bears a huge impact on human lives and rights.

The challenges posed by the migration lexicon are based on its internal stratification depending on factors such as: national, regional, and local differences, diversity in possible audiences (from institutional international stakeholders to aid workers and finally migrants themselves) and from the intrinsic non correspondence of different migration responses along with rapid changes in legislation for each recipient country taken into consideration.

At a macro level the migration lexicon can be organized into three large sub-set or domains differentiated by scope, and in some cases, partially overlapping:

1) An international or transnational level, here identified with the institution of the European Union’s regulations (legal and administrative/institutional and its migration approach principles, e.g., *Accordo di Cotonou, beneficiario di protezione internazionale, cittadino di un paese terzo, Convenzione di Dublino, discriminazione diretta, migrante di seconda generazione*). This domain is best described since it needs to be standardized at least for all EU languages (although not taking into consideration the languages of migrants which are seldom corresponding to any of the above-mentioned languages) [[1]](#footnote-1). This level, from a translation point of view is medium-high standardized in terminology and TE within EU languages, but there is no standardization for languages outside EU languages and in TE. Available glossaries for this level contain a selection of about 500 entries at most and are not derived using corpus-based procedures. Examples of different kinds including Italian language are [ European Migration Network 2018 , IATE 2018 , International Organization for Migration (IOM) 2019].

2) The second macro sub-set is national in scope, and it regards procedures, regulations, adaptations, and additions that are modified and proposed by each specific country administrative and general migration policies. Each country does in fact implement and define regulations and implementations in individual and not cross-nationally comparable ways. These regulations and implementing decrees are constantly changing depending on government direction and overturns and on public opinion stances (e.g., in Italy for example Security decrees that have deeply changed in time asylum typologies in the last few years, etc.: *Agenzia del demanio, ente territoriale, certificazione sanitaria, ufficio G.I.P., SPRAR, abuso della libera circolazione, associazioni del Terzo Settore, autonomie locali*, etc.). The second layer is country-specific presents low standardization of terminology and TE within EU languages, with partial overlapping with level 1 and no standardization for languages outside the EU.

3) A last sub-set of the lexicon relevant to migration management process concerns activities that are interlinked to aspects that migrants must face in their interactions with institutions for social security, health, education, administrative issues etc. (*carta di identità, stato civile, certificato di matrimonio, tessera sanitaria*, etc.).

The basic needs regarding glossary enrichment concern: methodology (corpus-based, corpus design and updating); inclusion of languages of migrants; inclusion of the overall stratification of domains pertaining migration not focusing only on institutional and regulative texts. Furthermore, the migration domain is affected by a widespread media usage of terms in a non-technical sense generating potentially dangerous ambiguity (*illegal and legal immigration, refugee, economic migrant*, etc.). For all these sub-strata there is an urgent need to develop strategies and tools that cover significant gaps both at international and national levels for the interest of multiple stakeholders (lexicographers, international and national policy makers, aid workers and cultural mediators and migrants themselves).

***1.2. The Language on the Fly resource (LoF)***

Language on the Fly is a prototype resource for building corpus-based glossaries and resources in the domain of migration. It is meant to be a full-fledged free online platform which aims to provide reliable and updated linguistic, lexicographical, and documentary information for orientation language in the first reception of migrants and asylum seekers upon arrival in an EU country. The prototype of the platform, to be released by the end 2021, and covers basic local, national, and international terminology in Italian (as source language), English (as lingua franca) and Arabic (as target language).[[2]](#footnote-2)

The Language on the Fly project is by design monodirectional, since it aims to consider especially the lexicon that is used in country specific national, regional, and local texts on migration and that often do not conform to international standards.[[3]](#footnote-3)

The goal of the project is to bring together different types of operational, scientific, and psychosocial skills in a way to produce a model of linguistic guidance that is both synthetic and characterized by precision and translation accuracy, and usable for different types of audiences.

# 2. LoF resource structure and corpus-based approach

***2.1. The structure and examples of the LoF cycle***

The LoF project is organized in three stages involving the three languages included. The phases of processing are organized in cycles to benefit from the empirical approach and to consider updates in relevant documents and procedure.

Diagram

Description automatically generated

Fig. 1. Monolingual SL resource map

As can be observed in Fig. 1 the first step in glossary building is the definition of a starting lemma list. The lemma list is further increased by progressive cycles that enlarge the corpora for SL. As can be seen the starting point is a set of lemmas already collected in previous general glossaries including Italian as SL or TL. The starting list is made by 564 lemmas. But the core of the lemma list extraction is corpus based. Corpora representing the three levels of stratifications have been used and constructed for resource and are constantly updated. At the moment of writing, corpora are composed the following way:

Table 1. Overview Corpora used or created for LoF (05/2021)

| **Corpus**  **Title** | **Area** | **Original** | **Occurrences** | **Text**  **Typologies** |
| --- | --- | --- | --- | --- |
| A.1 Corpus of International EU Migration (Italian) | EU | yes | ca 5,000,000 | Documentation from the European Parliament, The Council of Europe, The Court of Justice, The European Commission on migration issues |
| A.2 Corpus of International EU Migration (English) | EU | yes | ca 5,000,000 | Parallel corpus to Corpus A1 |
| A.3 Corpus of International EU Migration 2 (Italian) | EU | yes | ca 1,000,000 | Documents produced within the EU that have no official translation in English so conceived as a comparable corpus |
| A.4. EUR-Lex Italian 2/2016 | EU | no | 606,070,097 | The EUR-Lex Corpus is a multilingual corpus in all the official languages of the European Union. Thanks to the coverage of a vast area of subjects, the corpus is an excellent general-purpose resource for anyone looking for translation examples in many languages. |
| A.5 EUR-Lex English 2/2016 | EU | no | 629,722,593 | The EUR-Lex Corpus is a multilingual corpus in all the official languages of the European Union. Thanks to the coverage of a vast area of subjects, the corpus is an excellent general-purpose resource for anyone looking for translation examples in many languages. |
| A.6 Comparable Migration Corpus (Arabic) | EU/Int | yes | ca 1,000,000 | Documents and translations transnational in nature regarding migration in Arabic. |
| B.1 Corpus of National Migration 1 (Italian) | Italy | yes | ca 5,000,000 | Italian legislation and documentation of migration management specific to the Italian context |
| B.2 Comparable National Migration Corpus (Arabic) | Arab countries | yes | ca 5,000,000 | Documents regarding legislation and management of migration in Arab countries. |
| C.1 Corpus of Subsidiary National Migration 1 (Italian) | Italy | yes | ca 5,000,000 | Documentation for social security, health, education, administrative issues etc. |

The first lemma list of LoF was extracted from corpora A.1 and B.1 by using keyword (both corpus internal TfIdF and using reference corpora comparison) and collocation extraction measures (N-grams, association measures) and further manually selected. The first release of the LoF resource is (provisionally) composed of 2,094 entries (of which 1,558 multiwords, and 89 named entities).

Following the definition of the entry list a set of procedures are applied for the description of the SL entry properties:

1) Identification of relationships (type of, hypernyms, hyponyms, related items) among lemmas by using external reference resource tree system, in our case EUROVOC descriptors, subjects and Directory code provided for the EU document database;

2) Identification of formal variants and synonyms: using both corpus internal (intra-language) and multilingual parallel corpora (inter-language) methods [A.1-A2, A.4-A.5] and comparable corpora [A.6, B-2], focusing only on meanings pertaining to the migration domain;

3. Identification and association of collocations and idioms (using N-grams, association measures and work sketches);

4. Extraction of primary source definition for key concepts with relative source: for each term concordances from corpora led us to identify technical definitions of terms, that were added with the source to the glossary. A prototypical alternative definition in a controlled language was further provided.

4. Extraction of usage examples from corpora: following guidelines that avoid editing materials, examples with named entities or dates, and length requirements, examples were selected for each entry.

The result of these operations is converted in a relational database form that produces a monolingual resource of terminology that is the starting point for the processing for all target languages.

For pure exemplification the entry for the SL has the following structure see Fig. 2.

![Graphical user interface, text, application, email

Description automatically generated]()

Fig. 2 Example of Monolingual SL resource database

One of the main challenges remains the absence of a common migration framework that grant good quality translations that do not introduce potentially risky ambiguities. In this respect glosses are provided for translations in cases where the cultural and legislative background demands it (some examples will be provided in the following par. 2.2).

We will now give an overview about quantitative data on the resource on some of the languages as it has been collected until the day of writing, see Table 2.

Table 2. Overview of Entry list of LoF (05/2021)

| **Languages**  **(ISO codes)** | **# entries** | **# multiwords** | **variants** | **synonyms** |
| --- | --- | --- | --- | --- |
| IT | 2,094 | 1,558 | 455 | 970 |
| EN | 2,094 | 1,609 | 519 | 1,196 |
| AR | 1,009 | 764 | 0 | 22 |

***2.2. Challenges in using corpora for the Arabic version of the glossary***

The choice of Arabic as one of the languages for the pilot study has not been casual since Italy is a country that in time has witnessed different waves of migration from Arabic speaking countries, mainly Maghreb and Egypt, but more recently since 2011 has been a final or intermediate destination of many Syrian refugees. The challenge lies in translating terminologies that point out the cultural differences in the Arab communities, to cite some examples: *migrazione* (migration), *matrimonio fittizio* (*marriage of convenience*) and *adozione fittizia* (*adoption of convenience*) but also terminology referring to migration itself. The root for the Arabic word هجر means both to migrate and to abandon. In order to express the verb *to migrate, or to immigrate*, in Arabic we choose a morphological form which is هاجر which expresses a long duration of abandonment/ migration. While in Italian language we can use 3 verbs to express the idea of leaving one’s country (*migrare, emigrare, immigrare*), in Arabic language we use the nominal verb only to express the three cases followed by an adjective هجرة وافدة، هجرة مغادرة، هجرة. The previous distinction is reflected also on the word *migrant*, *immigrant* and *emigrant*. It is worth noting that in Italian the last two terms are used in the past participle form, while it is used in the present participle form in English. It is necessary to add an adjective to specify the direction to express the word emigrant (he/she who leaves his/her own country) مهاجر مغادر, while we add another adjective to express the idea of the immigrant, he/she who arrives, we say مهاجر وافد which changes perspective.

*Matrimonio fittizio* (*marriage of convenience*) is a recent phenomenon which is becoming commonly practiced within Arab communities in the European or Western context, while it is less common in the Arab countries. The closest translation would be زواج صوري (“*false marriage”*) or زواج المصلحة (“*marriage of interest”*), was thus translated as an “artificial marriage” or “marriage for benefit”.

Whereas *adozione fittizia* (*adoption of convenience*) In the case of Italian *adozione fittizia* (*adoption of convenience)*. The first translation results in an explanation rather than a short translation as it is necessary to provide context and an explanation to the term as it is not simply an adoption, and it means ‘’adoption for the sake of receiving a residence permit’’, in this case, the word ‘’adoption’’ is the usual concept of adoption. The second one is translated into تبني صوري fake adoption’’ which comes from صورة which means “an image/photo”. Whereas the third translation احتضان means “embracing” a child for the sake of receiving the residence permit. While the fourth translation means “legal fostering of a child” or literally “sponsorship”, the origin of the word comes from كفل which also means guaranteeing for example food and care for the child. This translation is the mostly common and used in the Arab world as the cultural and Islamic context of the word adoption تبني is not commonly used by Arab Muslims because the implications of the two words “adoption” تبني and “legal fostering” كفالة are significantly and legally different. Adoption تبني involves giving the child the father’s name and family name, and the same child can fully inherit from his father. While in the fourth translation كفالة, which is the legal fostering, means taking care of the child fully as if he/she’s a biological child, however, the child cannot inherit the father’s fortune and at the same time cannot take the full name of the father (name and family name), the child can be given the first name of the father or his last name, but not the full name. Christian Arab families can fully adopt children granting them their full names and those children can inherit. The closest translation of the world adoption in a European context is تبني in which a child can be named after his father and can inherit. It is worth pointing out to the fact that a child can still inherit from his parents even within kafala كفالة through leaving a will.

It is possible to identify two main critical areas of translation that pose challenges in the case of building glossaries involving EU languages and Arabic: on one side, there is the language use and connotations, this area includes the terminology or word-usage whether in Modern Standard Arabic or in different spoken dialects across the Arab world. Depending on the word’s root or verbal form, the meaning can change, or it can bear different or involving controversial connotations. In addition, there are cases in which the is a missing translation equivalent or multiple equivalents or even translations that refer only to narrower concepts that need to consider. On the other hand, the social and cultural differences are also a significant area to be taken into consideration, for the legislative backgrounds that differ even within different Arab countries and in addition to the different practices based on religious or cultural habits that are not share in host countries.

References

*1.European Migration Network* (2018), Asylum and Migration. Glossary 6.0. European Migration Network (EMN), Belgium.

*2.IATE* (2018), Interactive Terminology for Europe. European Union, Belgium.

*3.International Organization for Migration (IOM)* (2019), International Migration Law N°34 - Glossary on Migration. International Organization for Migration (IOM), Interactive Terminology for Europe.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Isabella Chiari**

Sapienza Università di Roma (Italy).

***E-mail:*** [***isabella.chiari@uniroma1.it***](mailto:isabella.chiari@uniroma1.it)

**Maha Bader**

University of Bergamo (Italy).

***E-mail: maha.bader@guest.unibg.it***

**Alma Salem**

Pontifical Institute for Arabic and Islamic Studies (Italy).

***E-mail:*** ***alma.salem@pisai.it***

**Luigi Squillante**

Sapienza University of Rome (Italy).

***E-mail: luigi.squillante@uniroma1.it***

1. An example of a multilingual glossary of migration terms at EU level is the [European Migration Network 2018] it contains contains translations of terms into 22 of the EU’s 24 languages (about 520 lemmas). While the online version of the Glossary is version 7.0 and was updated in July 2020 with the addition of 28 new terms reflecting the most recent European policy on migration and asylum. [↑](#footnote-ref-1)
2. Working versions now are being developed six EU languages (Italian, English, French, Spanish, Portuguese, German) and 10 non-EU languages (Arabic, Azerbaijani, Serbian, Pashto, Russian, Persian, Albanian, Turkish, Chinese, Norwegian). Priority is given to languages that are most represented as countries of origin of migration in Italy. [↑](#footnote-ref-2)
3. The prototype itself can nevertheless be applied to any source country (and language) as a methodology and procedure for collection and description of data. [↑](#footnote-ref-3)