



RAEL: Revista Electrónica de Lingüística Aplicada

Vol./Núm.: 23/1
Enero-diciembre 2024
Páginas: 1-18
Artículo recibido: 14/08/2024
Artículo aceptado: 21/11/2024
Artículo publicado: 31/01/2025
Url: <https://rael.aesla.org.es/index.php/RAEL/article/view/668>
DOI: <https://doi.org/10.58859/rael.v23i1.668>

Compiling a Corpus of User-Generated Content Units for the Detection of Social Problems

Creación de un corpus de contenido generado por los usuarios para la detección de problemas sociales

ROCÍO JIMÉNEZ BRIONES
UNIVERSIDAD AUTÓNOMA DE MADRID

Social media platforms like Facebook, X, and Instagram provide valuable information about daily and global social problems through their user-generated content units. Such platforms turn their users into *social sensors* capable of identifying problems such as violence against women or natural hazards. Within the field of crowdsensing, which aims to extract useful information from these social sensors, we introduce a proposal for identifying problems in ALLEGRO (<http://allegro.ucam.edu/>). Based on previous studies within this smart multimodal system, our research first explains the conceptual framework for addressing one of these problems, namely, elderly institutional abuse, within the text analysis module of ALLEGRO or DIAPASON. We also detail the methodology and challenges of compiling a subcorpus of tweets related to this problem. Such a specific subcorpus will contribute to ALLEGRO's comprehensive corpus of social problems, which is being built as a training set for deep learning models in text classification.

Keywords: *subcorpus; user-generated content; elderly institutional abuse; ALLEGRO, DIAPASON*

Plataformas como Facebook, X e Instagram proporcionan información muy valiosa sobre problemas cotidianos y globales gracias al contenido generado por sus usuarios, ya que estos se convierten en *sensores sociales* capaces de identificar problemas como la violencia contra las mujeres o los peligros naturales. Dentro del campo del *crowdsensing*, cuyo objetivo es extraer información útil de estos sensores sociales, presentamos una propuesta de identificación de problemas en ALLEGRO (<http://allegro.ucam.edu/>). Basándonos en estudios previos sobre este sistema, se explica el marco conceptual para abordar el problema del maltrato institucional a ancianos dentro del módulo de análisis de texto de ALLEGRO o DIAPASON. Se detalla también la metodología y los retos que plantea la compilación de un subcorpus de tuits para dicho problema social, el cual contribuirá al corpus integral de problemas de ALLEGRO que se está construyendo como conjunto de entrenamiento para modelos de aprendizaje profundo en la clasificación de textos.

Palabras clave: *subcorpus; contenido generado por usuarios; maltrato institucional a mayores; ALLEGRO; DIAPASON.*

Citar como: Jiménez Briones, R. (2024). Compiling a Corpus of User-Generated Content Units for the Detection of Social Problems. *RAEL: Revista Electrónica de Lingüística Aplicada*, 23, 1-18. <https://doi.org/10.58859/rael.v23i1.668>

1. INTRODUCTION

Social media users have lately become rich sources of information on those issues that concern them in their daily life activities or in a more global perspective. Through text data, photos, videos, audios, or a combination of them, millions of opinions, worries, complaints, etc. that affect people's quality of life (QoL) are shared, every day, in social networks like Facebook, X (formerly Twitter), Instagram, or LinkedIn. As Ghani, Hamid, Targio Hashem, and Ahmed (2019: 418) explain:¹

All the statuses, tweets, comments, posts, and reviews are the user-generated content. User-generated content is a type of data that typically refers to images, text, and videos. This content comes from regular people and not necessarily in a standard form [...]. All these data may incorporate the users' personal opinions, behaviors, and thoughts, which makes the task of extracting high-quality information from such data becoming increasingly important.

Such a gigantic amount of information (whether structured, semi-structure and/or unstructured) that can be conveyed in various modalities (i.e. textual, visual, sound, etc.) is known as big data, which, along with the advances in computer tools, has opened a door not only to understanding human behavior but also to carrying out applied research, within Artificial Intelligence (AI) and data analytics, into areas like "health, astronomy, social network and geoscience" (Ghani et al., 2019: 417).

Companies and governments are interested in storing all the massive data generated in the various social media applications because, among other things, it allows them to collect customer feedback and reviews about their products, detect political tendencies, and locate disease outbreaks, catastrophes, traffic jams, etc. (Bazzaz, Haghi, Mahdipour & Jameii, 2021: 9). Thus, the individuals that create this user-generated content (UGC) units in the social networks are considered active social sensors (Sakaki, Okazaki & Matsuo, 2013; Musto, Semeraro, Lops & De Gemmis, 2015; Arthur, Boulton, Shotton & Williams, 2018) able to identify a great myriad of problems around them such as violence against women, pollution, bullying, floods, etc.² This has resulted in a multidisciplinary field of study known as *social sensing* or *social-media crowdsensing* (Li, Wu, Wang, Cheng, Chen, Zhou & Ding, 2017; Wang, Szymanski, Abdelzahr, Ji, & Kaplan, 2019). Although these terms are used differently in the literature (An & Weber, 2015), we understand that "social sensing through social media involves the analysis of digital communications to detect real-world events or situations" (Periñán-Pascual, 2023b: 264).

In this context, a large inter-university team of computer scientists and linguists is cooperating in the design of **ALLEGRO** (*Adaptive muLti-domain sociaL-media sEnsinG fRame-wOrk*), a smart multimodal system (text, audio, and image) for the prediction of social problems (<http://allegro.ucam.edu/>). Thus, within the multidisciplinary field of crowdsensing, whose ultimate goal is to derive content from social sensors by employing computational means, and drawing upon the works of Periñán-Pascual (2023a, 2023b, 2024a, 2024b), Alameda Hernández (2024), Felices Lago (2024, 2025), Jiménez-Briones and Felices Lago (2024), and Ureña Gómez-Moreno (2024), among others, we address a proposal for the detection of social problems related to the domain of POPULATION within ALLEGRO. In particular, zooming in on the problem of elderly institutional abuse, this research outlines the conceptual framework adopted for tackling such a problem in DIAPASON (*unifieD hybrId ApProach to microtext Analysis in*

¹ Following Ghani et al. (2019: 418), social media, social media sites and social networks are used interchangeably in this paper.

² Periñán-Pascual (2023b) details the main features that define the notion of UGC in the field. This research sides with his proposal "where (a) users play the role of citizens, and (b) the content describes events or situations perceived as problems that disrupt their QoL" (Periñán-Pascual, 2023b: 264).

Social-media crOwdseNsing), the ALLEGRO's module devoted to text analysis (<http://allegro.ucam.edu/diapason/Index.html>). Additionally, the steps followed to compile a corpus of tweets related to this problem are also detailed with a view to using the resulting subcorpus as part of the gold standard for the automatic detection of social problems by the ALLEGRO system (Periñán-Pascual, 2024b).

The article is organized as follows. Section 2 sketches the goals of ALLEGRO, delving into the ontological approach followed to address, linguistically and conceptually, social problems in DIAPASON. Section 3 presents the case study of one social problem: elderly institutional abuse. Its ontological treatment (3.1), codification in the format of a problem schema (3.2), and the compilation of a subcorpus of UGC units dealing with said problem (3.3) are fully detailed, too. Finally, the conclusions are presented in Section 4.

2. AN OVERVIEW OF ALLEGRO AND DIAPASON

As introduced in Section 1, ALLEGRO is a project for the development of real-time crowdsensing applications that can accurately depict the state of society as envisaged by the collective intelligence of the social networking platforms' users (Periñán-Pascual, 2024b). It comprises two components: Data Analysis and Data Fusion (Periñán-Pascual, 2023a: 226). The former, in turn, is built upon three modules, one for each of the data required (i.e. text, audio, and image): DIAPASON (text analysis), SOUND (*Social-media sOUNd aNalysis moDule*), and ADAGIO (*SociAl meDia imAGe analysIs moDule*), respectively. Thus, when microtexts incorporate audio and/or images, this expanded knowledge is combined in Data Fusion “by rejecting irrelevant information, minimising redundancy, resolving inconsistencies, and completing missing information” (Periñán-Pascual, 2023a: 226). Both Data Analysis and Data Fusion use a multi-modal data repository and a knowledge base. In other words, the same ontology and formalism (i.e. problem schemas) are shared in the ALLEGRO modules to model the knowledge obtained from the UGC units and, accordingly, formalize the most relevant features of the social problems expressed therein.

It is worth stressing that the ALLEGRO system needs to be understood within the field of social sensing or, more specifically, social media crowdsensing, whose goal is to infer knowledge through computational models by collecting, analyzing and interpreting UGC items (Li et al., 2017; Wang et al., 2019). This line of research views social network users as sensors of the society in which they live, as the problems that surround them are widely shared in posts, tweets, microblogs, etc. on daily basis. Consequently, the research developed within ALLEGRO contributes to the notion of *Smart City*: “a city that manages, in an intelligent way, all its associated resources with the aim to enhance the quality of the services provided to citizens and to improve their quality of life” (Espinoza-Arias, Poveda-Villalón, García-Castro & Corcho, 2019 as cited by Periñán-Pascual, 2023b: 265). With this aim, UGC units are looked into to uncover any type of problem that can negatively affect a community. Our focus, however, extends beyond the quality of urban infrastructure and services provided to citizens (e.g. water/gas/electricity supplies, waste network, healthcare, etc.) to include the sociological dimension of the city, as reflected in people's concerns about violence, ageing, crime, and education, among others. In this way, our proposal presents a more comprehensive alternative to already-existing smart city applications, which tend to focus primarily on the former aspect.³

It turns crucial then to comprehend the natural language that social sensors use to express the issues that impact their QoL and well-being, both for AI and for natural language processing (NLP). Thus, we, as linguists within the ALLEGRO project, devote ourselves to the analysis

³ The interested reader can find a detailed account of the various frameworks proposed to explore smart cities in Periñán-Pascual (2023b).

of English and Spanish microtexts in UGC through DIAPASON, “a workbench of exploration and experimentation with UGC written in English or Spanish to automatically detect a variety of problem types by integrating natural language processing (NLP), machine and deep learning, and knowledge engineering techniques” (Periñán-Pascual, 2023b: 265). As one of the DIAPASON’s goals is to assist the NLP module to accurately classify the topics and recognize the keywords of the problems posted in UGC items, let us briefly comment on its ontology, which is structured into 4 levels: problem realms, problem dimensions, problem domains and problem types (Periñán-Pascual, 2023a, 2023b):

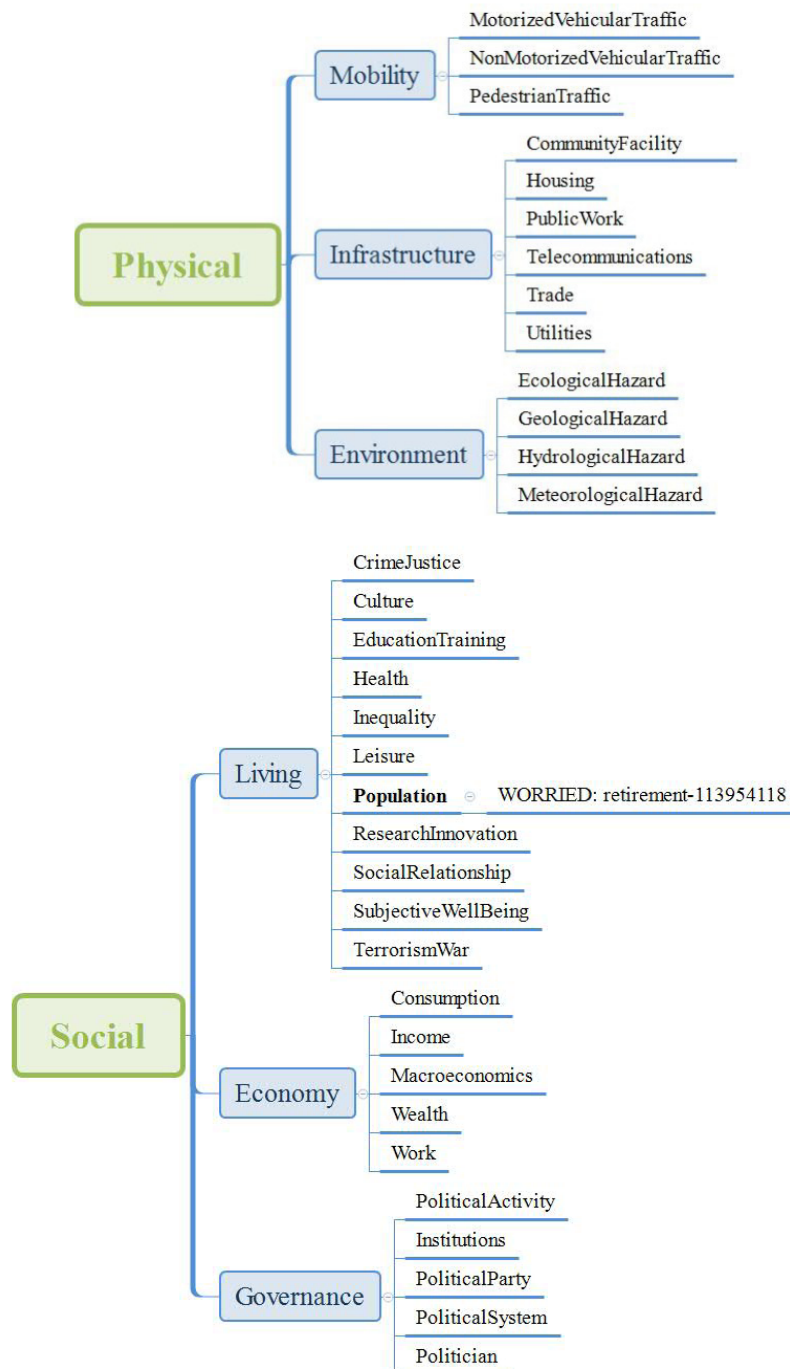


Figure 1. Architecture of the DIAPASON ontology

As Figure 1 displays, the upper ontological level comprises the PHYSICAL and Social problem realms, and six problem dimensions: MOBILITY, INFRASTRUCTURE, ENVIRONMENT, LIVING, ECONOMY, and GOVERNANCE.⁴ The intermediate ontological level includes the problem domains, such as COMMUNITYFACILITY or SUBJECTIVEWELLBEING, which store most of the areas about which citizens can complain in a community, whereas the lower ontological level refers to the specific types of problems that affect those who live in smart cities, using for that purpose a language-independent knowledge structure known as *problem schema*. For example, in the social domain POPULATION the problem *fear of retirement* is represented by means of the problem schema *WORRIED: retirement-113954118*. We turn now to explore the key role of problem schemas in DIAPASON.

As already mentioned, one of the central purposes of DIAPASON consists in automatically detecting the community problems that affect the citizens as they publish them in social media microtexts. To this end, this module first needs to have stored these problems in the ontology. Thus, the DIAPASON ontology conceptualizes the semantics of the community problems by means of problem schemas, which are the constructs employed to formalize such problems in the SOCIAL and PHYSICAL realms of the ontology. Below is the interface for editing said ontology and incorporating the various problem schemas:

DIAPASON Knowledge Organization System [log out]

Social Physical

Dimension: Living, Economy, Governance

Domain: Health, Inequality, Leisure, Population

Problem type: ElderlyNeglectCaretaker, ElderlyPhysicalAbuse, ElderlyPsychologicalAbuse, FearBecomingIllDisabled

New

ID + Name: #1107016

Description: Fear /anguish of becoming ill or disabled

Schema: WORRIED: ((elderly-107943870 | become-200149583) & (ill-302541302 | disabled-301019283))

Update Delete Validate

WordNet Entity

English Spanish

n v adj adv

Search

Figure 2. The DIAPASON Editor

⁴ The modeling of this ontological level stems from foundational studies on smart cities like Giffinger, Fertner, Kramar, Kalasek, Pichler-Milanovic, and Meijers (2007), Govada, Spruijt, and Rodgers (2017), and Appio, Lima, and Par (2019). A comprehensive account of how this level and the intermediate one were designed and implemented can be found in Perrián-Pascual (2023b, 2024a, 2024b). The lower level of the ontology is still being populated by the ALLEGRO linguists, although many of the community problems belonging to dimensions such as ENVIRONMENT and GOVERNANCE (Perrián-Pascual, 2024a, 2024b), LIVING (Alameda Hernández, 2024; Jiménez-Briones & Felices Lago, 2024; Ureña Gómez-Moreno, 2024), and ECONOMY (Felices Lago, 2024, 2025) have already been identified, listed and formalized.

The researchers modeling this type of knowledge in the ontology must either select or fill in the appropriate information in the DIAPASON Knowledge Organization System for each of the social problems already identified. By way of illustration, Figure 2 displays the case of the problem type *fear of becoming ill or disabled*, for which the appropriate dimension (LIVING) and domain (POPULATION) need to be selected in the interface. Likewise, an English short description of the problem must be typed in, along with its given name, which takes the shape of a list of running characters without spaces, using capital letters for the initial of each content word in the name: FEARBECOMINGILLDISABLED. Finally, the problem schema itself is provided: WORRIED: ((elderly-107943870 | become-200149583) & (ill-302541302 | disabled-301019283)). A word is needed to outline the two building blocks that make up DIAPASON problem schemas, that is, the illocutionary and locutionary components:

Table 1. Components of problem schemas

Illocutionary component	WORRIED
Locutionary component	((elderly-107943870 become-200149583) & (ill-302541302 disabled-301019283))

On the one hand, the illocutionary component expresses the non-propositional meaning of the schema, in this case, as Table 1 captures, through the communicative function WORRIED, which reflects the users’ negative perception of this issue and which will activate linguistic expressions, in the microtexts, such as *I am worried that []*, *I fear that []* or *What if []?*⁵ This component, which relies upon Searle’s (1969) Speech Act Theory, is optional as not all the community problems require the actual lexicalization of the users’ negative perceptions.⁶ On the other hand, the locutionary component is compulsory because it contains the propositional meaning of the schema, that is, the semantic content of the problem itself. To do so, WordNet synsets (Princeton University, 2010) are employed as the conceptual units of the description, i.e. *elderly-107943870*, *become-200149583*, *ill-302541302* and *disabled-301019283*, as well as several logical and conceptual operators to connect the selected synsets and reflect notions of addition (&), inclusive/exclusive disjunction (|, ^), quantity (M, P) and negation (N). It should be noticed that direct access to the WordNet database is granted by selecting the WordNet option in the Editor and typing in, in natural language, each of the key words of the problem under study (e.g. “ill”), which will list the appropriate synsets with their original numerical code. Thus, the problem schema shown in Table 1 should be interpreted as the sensor’s worry about ageing in relation to becoming sick and/or disabled.

After identifying, listing and formalizing as problem schemas the main community problems related to the domains established in Figure 1, the ALLEGRO linguists moved to the second stage in the process of ontological modelling in DIAPASON: the compilation of the subcorpora of microtexts from X. This is a crucial step in the ALLEGRO project because, to test how effective the DIAPASON module is in the detection of problems reflected in UGC units, as they are shared by citizens in the social media, these subcorpora will eventually result in a gold standard corpus for the training phase of the ALLEGRO algorithm, whose ultimate goal is to automatically sort out any microtext as either related or unrelated to a specific social problem. As Periñán-Pascual (2024b: 20) points out: “whereas the corpus is characterised by

⁵ Fernández-Martínez (2024) lists the set of communicative functions and their linguistic realizations that can be used to express the social sensors’ attitude towards a topic.

⁶ Although indirect speech acts are pervasive in language —cases in which a communicative function is present in a text but without a clear systematic linguistic correspondence (e.g. Holtgraves, 1994; Stefanowitsch, 2003; Pérez, 2013; Trott & Bergen, 2017, to name a few) —, we are not yet able to address them at this stage of the project.

linguistic heterogeneity, each subcorpus displays more homogeneity. As a result, this division of labour considerably reduces the time of corpus compilation.” Section 3 illustrates this phase with the social problem of elderly institutional abuse.

3. CASE STUDY: THE SOCIAL PROBLEM OF ELDERLY INSTITUTIONAL ABUSE IN DIAPASON

Section 2 has outlined the main goals of ALLEGRO, zooming in on DIAPASON, the module in charge of textual processing. Before addressing the compilation of a subcorpus of UGC units related to elderly institutional abuse (Section 3.3), the conceptual framework employed to deal with said problem is needed, in particular, a brief introduction about the ontological modeling of POPULATION (Section 3.1), within which the problem is included, as well as an explanation of the creation of its schema (Section 3.2).

3.1. Knowledge Modeling in POPULATION

According to the sociology literature (Parrillo, 2008; Bates & Ciment, 2013; Eitzen, Baca Zinn & Eitzen Smith, 2014; Marginean, 2014; Seccombe & Kornblum, 2020, among others), complaints about elderly institutional abuse belong in the field of aging, which, along with migration and tourism, comprise the three main areas of POPULATION as it is a domain that encompasses situations that can be understood as problematic due to the cumulative processes of people and the complaints of specific groups. It is one of the QoL indicators put forward and defined by OECD_iLibrary (*Organisation for Economic Co-operation and Development*) as “all nationals present in, or temporarily absent from a country, and aliens permanently settled in a country. This indicator shows the number of people that usually live in an area” (<https://www.oecd-ilibrary.org/>). As Figure 1 lays out, POPULATION in turn is included in the dimension LIVING, whose conceptual modelling has been exhaustively accounted for in Perrián-Pascual (2023b, 2024a, 2024b).

Within this context, our proposal for knowledge modeling in POPULATION has been realized through the identification, listing, and formalization of the citizens’ complaints and concerns in this domain. To this end, our methodology has followed three stages (Jiménez-Briones & Felices Lago, 2024). First, the specific area to be modeled was selected out of the three already-mentioned fields: aging, migration, and tourism. Second, a bibliographic consultation of the chosen topic in sociology manuals and encyclopedias, as well as in specialized academic journals such as *Australian Journal on Ageing*, *Australasian Journal on Ageing*, *International Journal of Geriatric Psychiatry*, and *International Journal of Environmental Research and Public Health* was carried out. Thanks to these sources, we listed the possible issues about which social media users would express their attitude: nursing homes, pensions, health, loneliness, etc. Finally, a new bibliographic consultation was accomplished to incorporate, into the catalog of problems, more specific objections or subjective perceptions of the community about the addressed domains. As documented in the OECD report *How’s Life? Measuring Well-Being* (2020), some well-being indicators for distinct population groups and across different countries are measured considering both objective and intrinsically subjective outcomes (cf. objective and subjective measures of Income and Wealth/Health/Safety or Subjective Well-Being). In our case, the more specific consultation of surveys, questionnaires and lists made to the community on what it considers a problem within the sub-theme of aging, such as *The Attitudes to Ageing Questionnaire* (Laidlaw, Power, Schmidt & the WHOQOL-OLD Group, 2007) or *The German Ageing Survey* (Klaus, Engstler, Mahne, Wolff, Simonson, Wurm & Tesch-Römer, 2017), helped us to refine the catalog obtained in the second methodological step and incorporate more particular and subjective problems such as abuse in nursing homes, age discrimination, fear of ending up in a nursing home or of becoming a widow or widower. As a result, Jiménez-Briones

and Felices Lago (2024) present the list of problem types developed to date (e.g. AGEISM, FEAR-NURSINGHOME, LOWPENSION, etc.), along with the formalization of this knowledge on aging in DIAPASON through the development of their corresponding problem schemas. Next section approaches the formalization of the problem type ELDERLYINSTITUTIONALABUSE.

3.2. Building the problem schema of elderly institutional abuse

As explained in Section 2, the construction and formalization of the problem catalog in DIAPASON is achieved by means of the interface functionality or Editor (see Figure 2). Furthermore, the researchers in charge of this knowledge modeling task followed the methodology discussed in Jiménez-Briones and Felices Lago (2024) and Felices-Lago (2025), illustrated in Figure 3 for elderly institutional abuse:

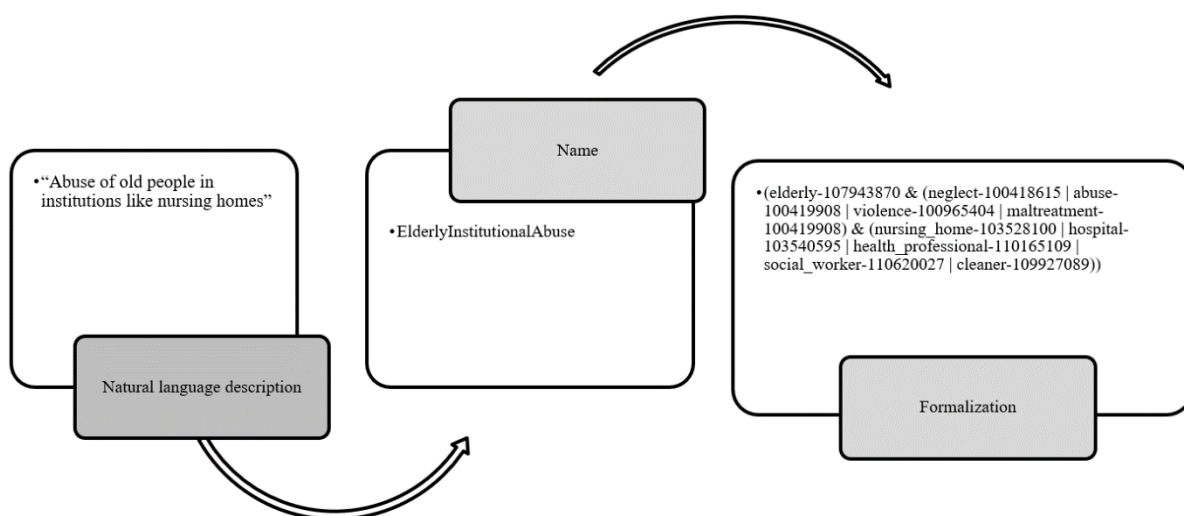


Figure 3. Problem schema methodology

Since, according to the literature (Rueda Estrada & Martín Martín, 2011; Walker, 2022), complaints about this type of abuse may involve different kinds of institutions (nursing homes, hospitals, clinics) and actors (physicians, social workers, nurses, etc.), the schema elaborated to define it, repeated in (1) for the sake of clarity, is certainly complex:

- (1) (elderly-107943870 & (neglect-100418615 | abuse-100419908 | violence-100965404 | maltreatment-100419908) & (nursing_home-103528100 | hospital-103540595 | health_professional-110165109 | social_worker-110620027 | cleaner-109927089))

First, as the synsets corresponding to *neglect*, *abuse*, *violence*, and/or *maltreatment* have negative connotations and already express a problematic situation, it was decided not to incorporate the illocutionary part into the schema. In these cases, it is better to opt for the minimal expression because, at this stage, the introduction of some communicative function in the problem schema (e.g. COMPLAIN) will force the machine to search, in the UGC units, for a linguistic expression that instantiates it (e.g. *This is unacceptable!*). That being so, the proposed schema in (1) only consists of the locutionary part, in which the following have been formalized:

- 1) Physical and psychological abuse of senior citizens. As a result, the synset *elderly-107943870* must always be present (&), followed by the synsets that conceptualize the type of abuse, that is, *neglect-100418615*, *abuse-100419908*, *violence-100965404* and/or (!) *maltreatment-100419908*.

- 2) The institutions that may be part of the abuse of the elderly. For this purpose, the synsets *nursing_home-103528100* and *hospital-103540595*, connected with the inclusive disjunction operator (*|*), have been incorporated.
- 3) People who could perpetrate the mistreatment. The presence of the inclusive disjunction between the synsets *health_professional-110165109*, *social_worker-110620027* and *cleaner-109927089* allows us to achieve this goal.

As inferred from above, the WordNet synsets turn crucial when codifying the conceptual content of each problem as a problem schema. WordNet (Princeton, 2010) is a comprehensive lexical database that organizes nouns, verbs, adjectives and adverbs into sets of cognitive synonyms, known as synsets. It also systematically documents the relationships between these synsets through semantic relations such as hyponymy, hypernymy, troponymy, holonymy, and meronymy. For this reason, if, for example, the hypernym *health_professional* is selected in the DIAPASON's knowledge organization system, no other items subordinate to it, such as *doctor*, *nurse*, *surgeon*, etc. will have to be housed in the problem schema as can be seen in the Related words box in Figure 4. In the same line, selecting the synsets *hospital-103540595*, *social_worker-110620027*, and *cleaner-109927089* already encompasses the different types of hospitals, social workers and cleaners.

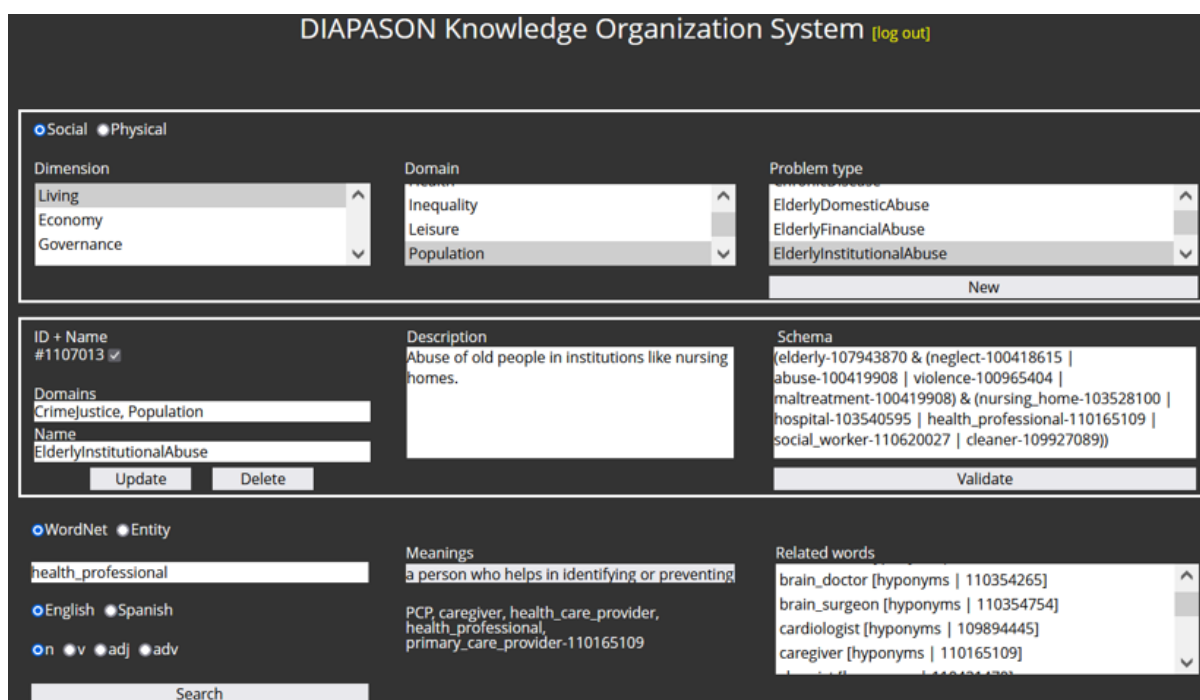


Figure 4. Elderly Institutional Abuse in DIAPASON

As made clear in Periñán-Pascual (2023a, 2023b), problem schemas are fundamental in the tasks of keyword recognition and topic categorization performed by DIAPASON. For the former, the synsets that are semantically connected to linguistic expressions in the microtexts are selected as keywords, whereas for the latter, problem schemas help to build a training set that DIAPASON will use to classify the UGC units. However, it is out of the scope of this paper to explore these two tasks in depth.

3.3. Compiling the subcorpus of UGC units on elderly institutional abuse

Employing elderly institutional abuse as a practical case, the two sections above have sketched the first stages for the automatic identification of social problems in the ALLEGRO system: a) choosing the problem domain from the DIAPASON ontology (e.g. POPULATION); b) identifying a set of socially significant problems within the domain based on the sociological literature available; c) selecting one problem and further consulting specific references (i.e. elderly institutional abuse); and, finally, d) building its problem schema with the DIAPASON Editor (Figure 4). This section adds a new and decisive step to the methodology: e) compiling a corpus of tweets on elderly institutional abuse. As mentioned in Section 2, the different subcorpora built for the DIAPASON problem schemas will contribute to the ALLEGRO gold standard corpus for the task of automatically sorting out UGC items as social problems or not. In consequence, the same attested methodology found in Periñán-Pascual (2024b), Alameda Hernández (2024) and Ureña Gómez-Moreno (2024) has been applied below.

3.3.1. Automatic extraction of microtexts

The first stage in the compilation of the subcorpus consists in gathering the necessary UGC units related to the problem under study. To this end, the application Twitter Searcher, developed by Periñán-Pascual (2024b: 20-21), was employed as it allowed us to access the Twitter API directly:

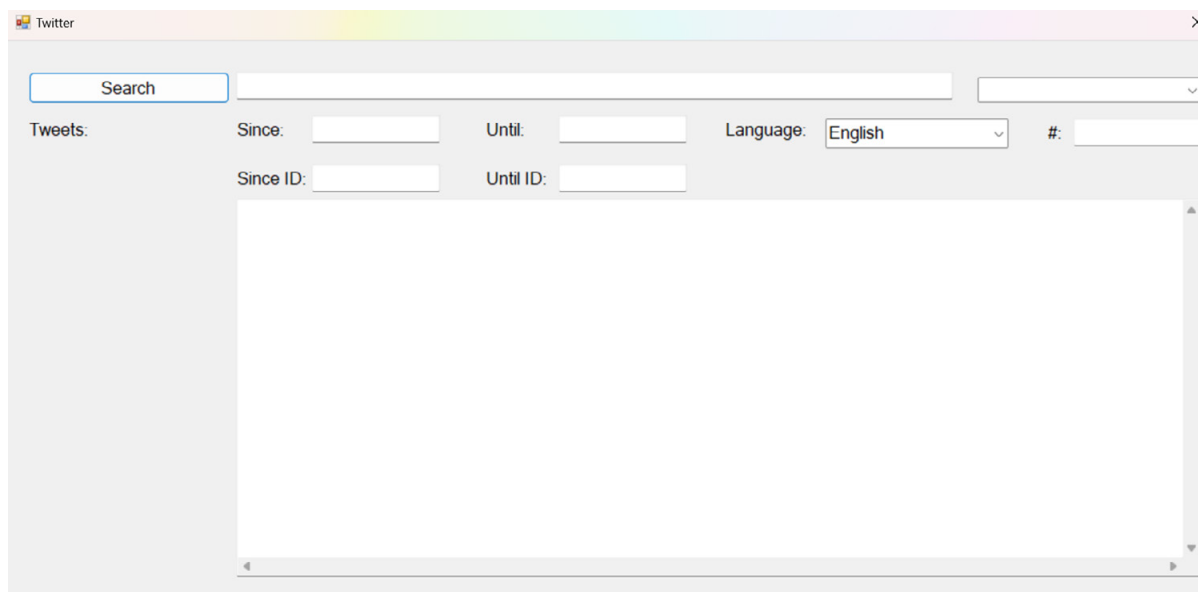


Figure 5. *Twitter Searcher* (Periñán-Pascual (2024b: 20-21))

The first query parameter in Twitter Searcher is the Search box, in which the keywords that will trigger the appropriate related tweets are typed in. As identified in the literature on building specialized corpora (Bowker & Pearson, 2002; Bathia, Sánchez Hernández & Pérez Paredes, 2011; Alayiaboozar & Hojjatpanah, 2022), this is a critical step when collecting data for a corpus because the choice of the searched words or seeds should strike a balance so that it neither biases the sample nor relies on introspection. Thus, drawing on Periñán-Pascual (2024b) and Ureña Gómez-Moreno (2024), whenever possible, keyword-based searches were avoided in favor of topic searches (e.g. words in hashtags, usernames, or named entities) so that the bias

towards using keywords from the very problem schemas for which we want to retrieve tweets could be minimized. Table 2 displays the topic words employed, from January to March 2023, to obtain tweets for our social problem:

Table 2. Search words for elderly institutional abuse

English	Spanish
elderly AND neglect AND institutions	tercera edad maltrato instituciones
elderly OR neglect OR institutions	tercera edad OR maltrato OR instituciones
seniors AND abuse AND nursing home	ancianos OR abuso OR residencia
seniors OR abuse OR nursing home	ancianos AND abuso AND residencia
seniors nursing home	ancianos residencias
seniors OR nursing home	ancianos OR residencias
seniors abuse	ancianos maltrato
seniors OR abuse	ancianos OR maltrato
elderly AND neglect	tercera edad OR maltrato
elderly OR neglect	tercera edad maltrato

In order to obtain a significant number of tweets, n-grams and various combinations with AND/OR were used, although some of the bigrams and trigrams were more productive than others. For instance, the word *institutions* was soon replaced with others because it retrieved tweets regarding universities and military or financial organizations.

Twitter Searcher also allowed us to set up the following parameters: a) Date: the time span for the searches or the tweet IDs could be established to avoid repeated microtexts; b) Language: English and Spanish tweets were retrieved; and c) Number: 500 tweets per query were introduced. Figure 6 provides a glimpse of one of the searches, whereas Figure 7 presents the results. Notice that, to filter out retweets and replies to other tweets, the instruction *filter:retweets-filter:replies* was included in the Search box.

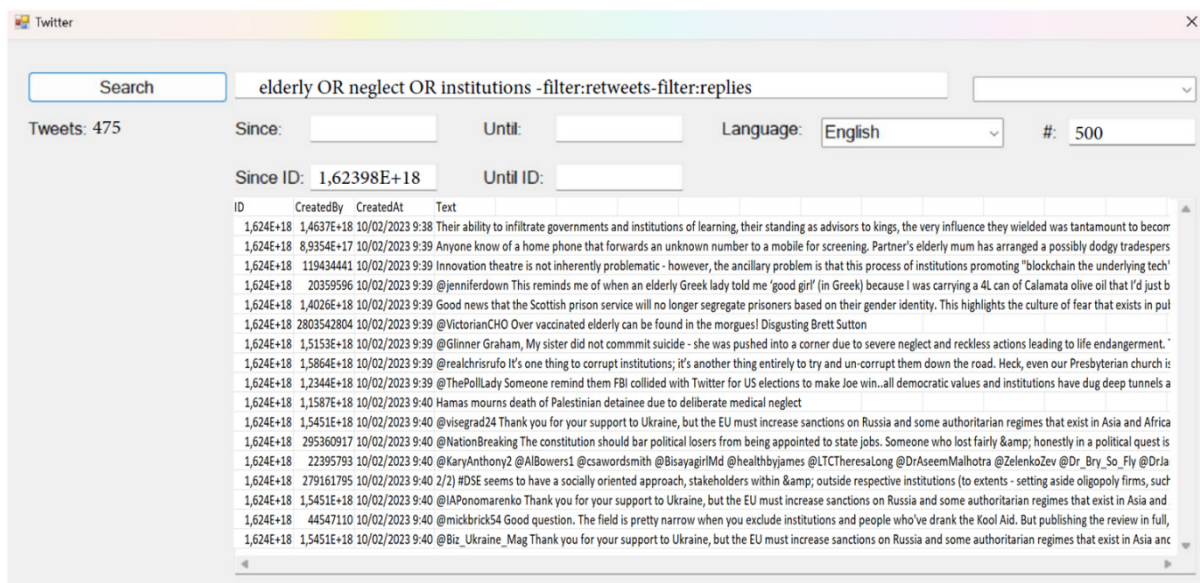


Figure 6. Twitter Searcher for elderly institutional abuse

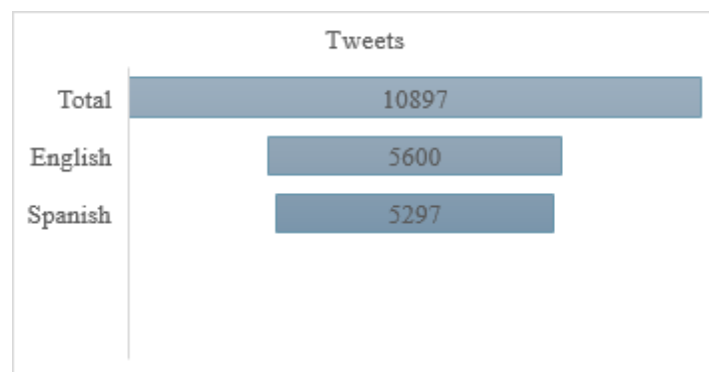



Figure 7. Number of tweets for elderly institutional abuse

Using Twitter Searcher on January 21, February 2, 10, and 28, 2023, we automatically obtained a total number of 10,897 tweets, out of which 5,600 were in English and 5,297 in Spanish. With such a vast number of tweets, we had to carry out a closer manual analysis to filter out those tweets that did not conform to our definition of social problem. Before continuing with the second step in the compilation of the subcorpus (see 3.2.2), a word is needed on the interpretation of social problems in ALLEGRO, which sides with the subjectivist approach found, among others, in Eitzen et al. (2014) and Best (1995, 2017), for which “social problems are what people view as social problems [...] No condition is a social problem until someone considers it a social problem” (Best, 1995: 4-5).⁷ In other words, a social problem comes about when society perceives it as a situation that harms or threatens the lives and well-being of some or many of the citizens and, accordingly, a solution needs to be found. Furthermore, such belief becomes a social problem only when it is collectively acknowledged and shared through social media such as X. As Periñán-Pascual states (2023b: 272), “the problem arises only when people make claims about it”. This way, such a significant number of tweets about elderly abuse indicates that this issue is viewed as a social problem by the community.

Based on this definition of social problem, the linguists within the ALLEGRO project agreed upon three criteria to identify and include UGC items that present social problems in each subcorpus: contemporariness, self-containment, and problem content. How each criterion is applied to elderly institutional abuse tweets is detailed below.

3.3.2. Manual filtering

The second step in the collection of the specialized corpus under study comprises manually annotating each tweet retrieved as 0 (off-topic related) or 1 (topic related). To filter in and out the microtexts that present UGC related to elderly institutional abuse, the three criteria mentioned above were followed. As for contemporariness, only tweets that make claims about institutionalized elder abuse for today’s society were selected. Example (2) illustrates such a case:

- (2)  Aged care worker, 56, charged with sexual assaults of elderly residents (25/01/2023)

As Alameda Hernández (2024) details, what citizens regard as problems varies throughout history, so there must exist a defined timeframe for identifying a specific social problem. Therefore, “the selection process excludes tweets that, although posted within the selected time span, deal with problems that have affected society in the past, even if the condition that motivated

⁷ In sociology, social problems have been addressed from two approaches: objectivism and subjectivism. The interested reader is referred to Periñán-Pascual (2023b) for a detailed explanation of each perspective.

the problem may still exist, but there is not a shared belief that perceives that as a problem at present” (Alameda Hernández, 2024: 71). Likewise, texts that deal with future or hypothetical problems like (3) were discarded:

- (3) Now u cannot protect yourself and an invalid. Now most likely she'll end up in a nursing home dead within a week from neglect, and God only knows what because services are overwhelmed in the area. And I bet you, Narcos buy the land. U 🤔 (09/02/2023)

The second criterion employed to manually filtering out off-topic microtexts is self-containment. As the subcorpus includes English and Spanish tweets, any fluent speaker of these languages must be able to grasp the content of the tweet and recognize it as addressing a social problem regardless of their cultural background or identity. For this criterion it is particularly relevant the use of names or abbreviations that refer to individuals, locations, or organizations that are understood only on a local or national level:

- (4) Casi 8k de ancianos de IDA, no Podemos, IDA, gracias a sus protocolos de exclusión y la negativa de medicalizar residencias, pero tu media neurona no puede procesar tanta información 🤔😅 (28/02/2023)

Example (4) was discarded because it includes the Spanish acronym *IDA*, which stands for Isabel Díaz Ayuso, President of the Community of Madrid, who many hold responsible for the elderly that died from COVID in nursing homes during the lockdown in Madrid. However, tweets that contain abbreviations that have made their way into the language (i.e. lol, omg, b4, ICU, ER, etc.) or are field-specific like CNAs (*certified nursing assistants*) or CdS (Spanish for *health center* or *centro de salud*) have been included in the corpus.

Finally, the filtering process also involves sorting out microtexts that, despite mentioning a problem theme, lack substantive problem-related content. In other words, the third criterion helps “to differentiate and filter the tweets that deal with the subjective belief as a problem affecting society and discard those tweets that are merely referential and deal with the social problem as a topic” (Alameda Hernández, 2024: 72). In our case, only UGC items like (5), which reflects abuse in elder care institutions, were included, discarding microtexts like (6), which announces some service for reporting cases of abuse in nursing homes. Our subcorpus, thus, collects tweets that showcase the social media users’ subjective perceptions and complaints about elderly institutional abuse.

- (5) The way seniors light up once you mention their children, its like the only thing they feel is worth remembering from life. Elderly abuse in care homes is bad. 🤖 The one's treated like crap the most is the one's that don't have family to advocate for them yes even the rich ones (28/02/2023).
- (6) Sadly, it is all too common for seniors living in nursing homes to experience abuse or neglect. Call 📞 910-839-7433 or visit 🏠 #nursinghome #nursinghomeneglect #nursinghomenegligence (03/02/2023)

3.3.3. Upcoming steps

Once the appropriate UGC units on elderly institutional abuse have been selected, the following steps need to be undertaken in the compilation of the subcorpus and in the completion of the ALLEGRO system. First, the reliability of the manual filtering explained above will be tested with another annotator. To this end, the Kappa index will be employed to evaluate the accuracy

and consistency of our tweet annotations. Second, other subcorpora need to be compiled for the rest of social problems belonging to the different dimensions of DIAPASON (see Figure 1). Thus, after the collection of the distinct subcorpora for the problem schemas in DIAPASON, linguistic studies could be pursued (e.g. word frequency lists, collocations, concordances, etc.) to research, among others, positive and negative attitudes of the social media users on topics such as institutionalized elder abuse, violence against women, unemployment, and so on. To do so, TexMiLab is the software designed by Perrián-Pascual (2024c) as a linguistic laboratory that incorporates tools for corpus building, text preprocessing, and text analysis.

It is worth highlighting that, in the light of the corpus data obtained for elderly institutional abuse, some of the already codified problem schemas for aging had to be redefined. For instance, the problem schemas ELDERLYABUSE, ELDERLYDOMESTICABUSE and ELDERLYFINANCIAL-ABUSE had to be reformalized differently, and even discarded in the case of the first schema, so that they did not overlap with ELDERLYINSTITUTIONALABUSE as finally defined in (1). Furthermore, while manually filtering out the tweets whose content was or was not related to elderly institutional abuse, we could identify UGC units for other social problems: ageism, fear of digital divide, elderly financial/domestic abuse, etc.

4. CONCLUSIONS

In the era of digital social media like X, Facebook, Instagram, etc., users have emerged as significant sources of immediate and authentic information through the content they post online or UGC. The identification and analysis of UGC can be highly valuable, particularly when such content addresses issues impacting society like elderly institutional abuse since exploring UGC can allow relevant institutions to take appropriate measures to address such specific social challenges.

In this context, the present study has focused on the compilation of a subcorpus of text messages from the social media platform X concerning elderly institutional abuse. This subcorpus has been incorporated in ALLEGRO, a multimodal intelligent system for the real-time identification and analysis of community problems published in social media platforms. After describing the ALLEGRO architecture, in particular its text-processing module or DIAPASON, as the corpus under analysis comprises written UGC, the methodological stages established for the completion of the intelligent system, as well as the challenges that lie ahead in the project, have been illustrated with the practical case of the problem of elderly institutional abuse. As one of the purposes of ALLEGRO is to build a medium-sized corpus of social problems that can serve as a training set for deep language learning models in text classification (Perrián-Pascual, 2024b), our proposed subcorpus contributes to the creation of a gold standard corpus for the automatic detection of social problems by this multimodal intelligent system.

ACKNOWLEDGEMENTS

This publication is part of the We-Collab project 2021-1-HR01-KA220-HED-000027562, co-funded by the Erasmus+ Programme and the R&D&i project PID2020-112827GB-I00, funded by MICIU/AEI/10.13039/501100011033.

REFERENCES

- Bazzaz, A. S., Haghi, K. M., Mahdipour, E., & Jameii, S. M. (2021). Big data analytics meets social media: A systematic review of techniques, open issues, and future directions. *Telematics and Informatics*, 57, 101517. <https://doi.org/10.1016/j.tele.2020.101517>
- Alameda Hernández, Á. (2024). Social media detection of texts on the social problem of violence against women within the multimodal intelligent system ALLEGRO. In R. Jiménez-Briones & A. Corral Esteban (Eds.), *Approaches to Knowledge Representation and Language* (pp. 63-75). Granada: Comares.
- Alayiaboozar, E., & Hojjatpanah, A. A. (2022). Steps for creating two Persian specialized corpora. *International Journal of Information Science and Management (IJISM)*, 20(4), 231-243.
- An, J., & Weber, I. (2015). Whom should we sense in “social sensing” - analysing which users work best for social media now-casting. *EPJ Data Science*, 4(22), 1-22. <https://doi.org/10.1140/epjds/s13688-015-0058-9>
- Appio F. P., Lima, M., & Paroutis, S. (2019). Understanding Smart Cities: Innovation ecosystems, technological advancements, and societal challenges. *Technological Forecasting & Social Change*, 142, 1-14. <https://doi.org/10.1016/j.techfore.2018.12.018>
- Arthur, R., Boulton, C. A., Shotton, H., & Williams, H. T. P. (2018). Social sensing of floods in the UK. *PLoS ONE*, 13(1), e0189327. <https://doi.org/10.1371/journal.pone.0189327>
- Bates, C. G., & Ciment, J. (Eds.) (2013). *Global Social Issues: An Encyclopedia*. New York: Sharpe Reference.
- Bathia, V., Sánchez Hernández, P., & Pérez Paredes, P. (2011) Specialized languages: Corpora, meta-analyses and applications. *Researching Specialized languages*, 47(1), 1-8. <https://doi.org/10.1075/sci.47.02bha>
- Best, J. (1995). Typification and social problems construction. In J. Best (Ed.), *Images of Issues: Typifying Contemporary Social Problems* (pp. 1-10). London: Routledge.
- Best, J. (2017). *Social Problems* (3rd ed.). New York City: W.W. Norton & Company.
- Bowker, L., & Pearson, J. (2002) *Working with Specialized Language: A Practical Guide to Using Corpora*. London and New York: Routledge. <https://doi.org/10.4324/9780203469255>
- Eitzen, S., Baca Zinn, M., & Eitzen Smith, K. (2014). *Social Problems* (13th ed.) Boston: Pearson.
- Espinoza-Arias, P., Poveda-Villalón, M., García-Castro, R., & Corcho, O. (2019). Ontological representation of smart city data: From devices to cities. *Applied Sciences*, 9(1), 1-23. <https://doi.org/10.3390/app9010032>
- Felices Lago, Á. (2024). Description of social problems by means of schemas related to the Income domain in the DIAPASON platform. In R. Jiménez-Briones & A. Corral Esteban (Eds.), *Approaches to Knowledge Representation and Language* (pp. 27-43). Granada: Comares.

Felices Lago, Á. (2025). Towards the characterization of WORK problem schemas in the DIAPASON ontology. *Sintagma* 37.

Fernández-Martínez, N. J. (2024). Exploring the creation of synthetic corpora of negative communicative functions for the task of communicative function identification. In R. Jiménez-Briones & A. Corral Esteban (Eds.), *Approaches to Knowledge Representation and Language* (pp. 77-93). Granada: Comares.

Gething, L. (1994). Health professional attitudes towards ageing and older people: Preliminary report of the reactions to Ageing Questionnaire. *Australian Journal on Ageing*, 13(2), 77-81. <https://doi.org/10.1111/j.1741-6612.1994.tb00646.x>

Gething, L., Fethney, J., McKee, K., Goff, M., Churchward, M., & Matthews, S. (2002). Knowledge, stereotyping and attitudes towards self-ageing. *Australasian Journal on Ageing*, 2(2), 74-79. <https://doi.org/10.1111/j.1741-6612.2002.tb00421.x>

Ghani, N. A., Hamid, S., Targio Hashem, I. A., & Ahmed, E. (2019). Social media big data analytics: A survey. *Computers in Human Behavior*, 101, 417-428. <https://doi.org/10.1016/j.chb.2018.08.039>

Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanovic, N., & Meijers, E. (2007). *Smart cities - Ranking of European medium-sized cities*. Retrieved from http://www.smart-cities.eu/download/smart_cities_final_report.pdf.

Govada, S. S., Spruijt, W., & Rodgers, T. (2017). Smart city concept and framework. In T. M. V. Kumar (Ed.), *Smart Economy in Smart Cities. Advances in 21st Century Human Settlements* (pp. 187-198). New York City: Springer. https://doi.org/10.1007/978-981-10-1610-3_7

Holtgraves, T. (1994). Communication in context: Effects of speaker status on the comprehension of indirect requests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(5), 1205-1218. <https://doi.org/10.1037//0278-7393.20.5.1205>

Jiménez-Briones, R., & Felices Lago, Á. (2024). La formalización del conocimiento en DIAPASON a través de una muestra de problemas poblacionales y macroeconómicos. In F. Olmo-Cazeveille (Ed.), *Investigación Lingüística en Entornos Digitales* (pp. 187-216). Granada: Tirant Lo Blanch.

Klaus, D., Engstler, H., Mahne, K., Wolff, J. K., Simonson, J., Wurm, S., & Tesch-Römer, C. (2017). Cohort Profile: The German Ageing Survey (DEAS). *International Journal of Epidemiology*, 46(4), 1105-1105. <https://doi.org/10.1093/ije/dyw326>

Laidlaw, K., Power, M. J., Schmidt, S., & the WHOQOL-OLD Group (2007). The attitudes to ageing questionnaire (AAQ): Development and psychometric properties. *International Journal of Geriatric Psychiatry*, 22, 367-379. <https://doi.org/10.1002/gps.1683>

Lee, H. J., Lee, D. K., & Song, W. (2019). Relationships between social capital, social capital satisfaction, self-esteem, and depression among elderly urban residents: Analysis of secondary survey data. *International Journal of Environmental Research and Public Health*, 16, 1445, 2-13. <https://doi.org/10.3390/ijerph16081445>

Li, W., Wu, W., Wang, H., Cheng, X., Chen, H., Zhou, Z., & Ding, R. (2017). Crowd intelligence in AI 2.0 era. *Frontiers of Information Technology & Electronic Engineering*, 18, 15-43. <https://doi.org/10.1016/j.intell.2017.04.004>

Marginean, I. (2014). Quality of Life Diagnosis (QoLD). In A. C. Michalos (Ed.), *Encyclopedia of Quality of Life and Well-Being Research* (pp. 5333-5339). Dordrecht: Springer. https://doi.org/10.1007/978-94-007-0753-5_2358

Musto, C., Semeraro, G., Lops, P., & De Gemmis, M. (2015). CrowdPulse: A framework for real-time semantic analysis of social streams. *Information Systems*, 54, 127-146. <https://doi.org/10.1016/j.is.2015.06.007>

OECD iLibrary (Organisation for Economic Co-operation and Development). Retrieved from <https://www.oecd-ilibrary.org/>.

OECD (2020). *How's Life? 2020: Measuring Well-Being*. Paris: OECD Publishing. Retrieved from https://www.oecd-ilibrary.org/economics/how-s-life/volume-/issue-_9870c393-en. <https://doi.org/10.1787/9870c393-en>

OECD (2022). "Population" (indicator). <https://doi.org/10.1787/d434f82b-en>

Parrillo, V. N. (Ed.) (2008). *Encyclopedia of Social Problems*. Los Angeles: Sage. <https://doi.org/10.4135/9781412963930>

Pérez, L. (2013). Illocutionary constructions: (Multiple source)-in-target metonymies, illocutionary ICMs, and specification link. *Language & Communication*, 33(2), 128-149. <https://doi.org/10.1016/j.langcom.2013.02.001>

Periñán-Pascual, C. (2023a). Exploring user-generated content to detect community problems: The ontological model of ALLEGRO. In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023) - V. 2* (pp. 224-230). KEOD. <https://doi.org/10.5220/0012203300003598>

Periñán-Pascual, C. (2023b). From Smart City to Smart Society: A quality-of-life ontological model for problem detection from user-generated content. *Applied Ontology*, 18(3), 263-306. <https://doi.org/10.3233/AO-230281>

Periñán-Pascual, C. (2024a). Modelización de las quejas de los ciudadanos como artefactos digitales culturales: DIAPASON. In F. Olmo-Cazeville (Ed.), *Investigación Lingüística en Entornos Digitales* (pp. 129-156). Valencia: Tirant Lo Blanch.

Periñán-Pascual, C. (2024b). Exploring problems through social media: The case of Beach Quality. In R. Jiménez-Briones & A. Corral Esteban (Eds.), *Approaches to Knowledge Representation and Language* (pp. 11-26). Granada: Comares.

Periñán-Pascual, C. (2024c). *Minería de textos para investigadores lingüistas*. Valencia: Tirant Lo Blanch.

Princeton University (2010). *About WordNet*. Retrieved from <https://wordnet.princeton.edu/>.

Rueda Estrada, J. D., & Martín Martín, F. J. (2011). El maltrato a personas mayores. Instrumentos para la detección del maltrato institucional. *Alternativas*, 18, 7-33. <https://doi.org/10.14198/ALTERN2011.18.01>

Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, 25(4), 919-931. <https://doi.org/10.1109/TKDE.2012.29>

Searle, J.R. (1969). *Speech Acts*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139173438>

Secombe, K., & Kornblum, W. (2020). *Social Problems* (16th ed.). Boston: Pearson.

Stefanowitsch, A. (2003). A construction-based approach to indirect speech acts. In K.U. Panther & L. Thornburg (Eds.), *Metonymy and Pragmatic Inferencing* (pp. 105-26). Amsterdam/Philadelphia: John Benjamins. <https://doi.org/10.1075/pbns.113.09ste>

Trott, S. and Bergen, B. (2017). A theoretical model of indirect request comprehension. *Proceedings of the AAAI Fall Symposium Series on Artificial Intelligence for Human-Robot Interaction* (pp. 129-132). Arlington, VA.

Ureña Gómez-Moreno, P. (2024) Integrating corpus methodology into the construction of an intelligent crowdsensing system. In R. Jiménez-Briones & A. Corral Esteban (Eds.), *Approaches to Knowledge Representation and Language* (pp. 45-62). Granada: Comares.

Walker, R. W. (2022). *Elderly Financial Abuse in New Zealand: Is the Law Sufficient?* (Master's dissertation). University of Canterbury. Retrieved from <https://ir.canterbury.ac.nz/items/5f894749-d5d1-4f47-a6c0-511e9bd299b7>.

Wang, D., Szymanski, B. K., Abdelzaher, T., Ji, H., & Kaplan, L. (2019). The age of social sensing. *IEEE Computer*, 52(1), 36-45. <https://doi.org/10.1109/MC.2018.2890173>