EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-11 September 2021

Ramada Plaza Thraki
Alexandroupolis, Greece

www.euralex2020.gr

Proceedings Book
Volume 1

Edited by Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-1-5

Edited by: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2020 Edition

Inventory of New Romanian Lexemes and Meanings Attested on the Internet

Barbu A.M., Lupu I., Stoica-Dinu O., Teleoacă D.L., Toroipan T.

Institute of Linguistics "Iorgu Iordan – Al. Rosetti", Romanian Academy, Bucharest, Romania

Abstract

This article presents a project that monitors the new lexemes and meanings attested on the Internet for the Romanian language and records them in a descriptive dictionary. This project tries to capture the dynamics of the language in the smallest details (e.g. the loan adaptation process) and to update the lexicographic inventory. The Internet is the best source for this purpose. The article begins with the definition of the term *new word* in the sense of this project, and with the characterization of a descriptive dictionary compared to a normative one. Then the method of selecting new words is described which is of random type, i.e. the words are selected by lexicographers from everyday life, without going through a predetermined volume of texts and are registered provided they have 10 attestations from different sources on the Internet. The article also provides a description of some technical aspects and the structure of the dictionary entries. Some solutions to the problems encountered, the first results and how to continue the project are also discussed.

Keywords: new words; Internet; descriptive dictionary; lemma variants; random selection

1 Introduction

The main goal of this paper is to present an overview of the project *Inventar de cuvinte și sensuri noi atestate în mediul online* 'Inventory of new Romanian lexemes and meanings attested on the Internet', henceforth ICSO, carried out at the Institute of Linguistics in Bucharest. The idea of building the ICSO was born from a previous project, coordinated by a software company, that had as objective the digitization of the Romanian lexicon in literal and phonetic form and with syllable separation (Diaconescu 2015). The digitally processed lexicon was extracted from three electronic explanatory dictionaries: DEX, MDA and DN. At one point, this lexicon was confronted with the words found in a pretty large electronic corpus of books and newspapers, belonging to the coordinating software company, and it was discovered that a relatively big number of common words in the corpus were not found in the digitized lexicon. This result showed the extent to which dictionaries had to be updated. Therefore, the team of the Institute of Linguistics participating in that project decided to inventory and gloss words that are not registered in the main general dictionaries of the Romanian language (which constitute an exclusion corpus – described below), extending the research from a particular corpus to the Romanian language used in the online environment, that is, generally, on the Internet. The preparatory work of ICSO started in 2017 and in 2019 the first volume of about 1400 entries was ready for publication (Barbu et al. 2020). ICSO is currently continuing with the second volume, with slightly changed lexicographic rules (more description-oriented) and as a database.

The project described here aims to monitor the Romanian language especially in the lexical aspect. It tries to update and keep up with the lexical explosion in the Romanian language produced by globalization, advanced technology and increasing access to information through digitization.

This lexical explosion is nowhere better seen than on the Internet, which gathers large volumes of highly diverse texts. "Legal, religious, literary, scientific, journalistic, and other texts will all be found there, just as they would in their non-electronic form", says David Crystal (2001: 31), who devotes an extensive study to the particularities of the language used on Web. Actually, using Internet as a corpus for linguistic (including lexicographic) research has already been set by linguists such as Fujii & Ishikawa (2000), Kilgarriff (2001), Grefenstette (2002) or Fuertes-Olivera (2012), among others. In addition to the benefits already mentioned in the literature, using the Internet as a lexicographic source presents some important features, from our point of view, such as:

- increased access to familiar language and internet slang on forums, comments, social media, etc.;
- access to the multitude of categories of commercial products from online stores, that deserve to be defined for the general public;
- documentation of the adaptation stages of some loanwords, by recording their graphical or morphological variants (e.g. smartphone also circulates in Romanian as smartfone and smartfon, with the plural smartphone-uri, smartfon-uri and smartfoan-e);
- greater opening of specialized terms to the general public.

The main purpose of this project is to build a descriptive new-lexemes inventory, which addresses the following aspects:

- to provide primary material for the explanatory dictionaries which decide what entered and what did not enter the language;
- to offer lexicographic support to the general public for as many terms as it can access on the Internet and not found in the published explanatory dictionaries;
- to provide a record of the words circulating at a given time in the language, even for a very short period, in support of linguists and other specialists from a distant future. Note that some new terms can reflect various language external

events. For instance, the actual coronavirus crisis has already coined or (re)vitalized in Romanian at least twenty candidates (e.g. *coronavirus, coronacriză* 'coronavirus crisis', *a carantina* 'to carantinate', *coronabonduri* 'corona bonds', *izoletă* 'isolation stretcher', etc.), who are likely to die sooner or later after this crisis is gone, but they could testify over the years for this social event;

- to offer a larger lexical resource for natural language processing (NLP), in a variety of tasks, such as automatic extraction of new words, marketing tasks and sentiment analysis, etc.

In order to make the design of ICSO and the working methodology clearer, in the next section we will define what we mean by new words (or new lexemes) and by what features a descriptive dictionary as ICSO differs from a normative one. The paper continues with a section describing the working method and the entry structure. The following section presents some aspects regarding the inventoried lexemes and the problems encountered in the construction of ICSO. The final section is dedicated to conclusions and further work.

2 Normative versus Descriptive Dictionary

In our opinion, the normative (or prescriptive) dictionary that includes new words is the result of a complex and rigorous *selection* process.

First of all, a *new word* is assimilated to the concept of *neologism*, defined for the first time by Zgusta (1971: 179): "neologism is a term which can refer to any new lexical unit, the novelty of which is still felt". Over time, this definition has been enriched with a series of requirements that constitute many criteria for the selection of words considered as neological. Thus, Cabré (1993: 445) mentions four parameters to determine the neological character of a lexical unit:

a. diachrony – a unit is neological if it has appeared in a recent period;
b. lexicography – if it does not appear in dictionaries;
c. systematic instability – if it shows signs of formal instability (morphological, graphical, phonetic) or semantic;
d. psychology – if speakers perceive it as a new unit.

Recent literature has paid more and more attention to the selection criteria for neologisms to be introduced in general dictionaries, sometimes following different lexicographic traditions (O'Donovan and O'Neill 2008, Adelstein and Freixa 2013, Sánchez Manzanares 2013, and others). By far the most adopted criteria are the frequency and the stability of new words in the language. Frequency refers to the dispersion in use (although some specialized terms are accepted), and stability refers to the period of time during which the new words are used, excluding (possible) ephemeral units. Another criterion is that of the neological feeling, which refers to the perception of the speakers about a word's novelty. This criterion excludes analyzable words (with a transparent meaning) from some dictionaries, but requires the introduction of less frequent words, in others (cf. Freixa & Torner 2019). However, the criterion with the highest prescriptive load is the denominative necessity. This refers to the exclusion of new words, often loanwords, which already have a correspondent in the target language.

Despite all these selection criteria, there are no purely normative dictionaries today. Words in nonstandard registers have special use marks (nonstandard, offending, regionalism etc.) that can be seen as prescriptive advices. However these words have entered the dictionary. If they were missing, one would not know whether a word does not exist in the general dictionary because it does not belong to the standard language or because it was accidentally omitted or because the general dictionary is not yet updated. As Rafel (2007: 20) claims, it has even been considered that the real limits between one type of dictionary and the other are not entirely precise.[1] However, a descriptive dictionary still differs by several essential features from one that is not purely prescriptive.

As stated in the literature, "a descriptive dictionary is one that attempts to describe how a word is used, while a prescriptive dictionary is one that prescribes how a word should be used (Naparsteck 2005: 28). In other words, a descriptive dictionary "aims to give a real and complete definition of each lexical item, without any restrictions based on prescriptive criteria." (Rafel & Soler 2016: 443). However, the descriptive character is understood differently in the specialized literature. One meaning is the one that refers to the way definitions are elaborated. For example, descriptive dictionaries could be called those that use, as definitions, detailed semantic descriptions and paraphrased meaning (in the usual sense of monolingual dictionaries) (cf. Gouws & Prinsloo 2005: 48-49). Another meaning refers to how words are defined based on their use. For instance, Collins COBILD formulates definitions in this way, e.g. the verb *condemn*: "If you *condemn* something, you say that it is very bad and unacceptable". Besides, this dictionary gives the syntactic pattern(s) in use for each word.[2] It should be added that descriptive dictionaries rely heavily on data collected from very large corpora. This is the case with dictionaries such as the Oxford English Dictionary and Collins COBILD.

In our opinion, a descriptive dictionary, dedicated to neologisms, closely monitors the language, especially pursuing lexical creativity and the way in which speakers use the available linguistic means. In these conditions, the descriptive dictionary is much less or not at all selective. It should not apply any of the above criteria and, thus, nonce words could find their place in such a dictionary. Furthermore, the systematic instability, which Cabré (1993: 445) spoke of, should also be reflected here, given that only a normative approach can establish a standard (morphological, graphical or phonetic) variant of several in use. The use marks (nonstandard, offending, regionalism etc.) have no prescriptive role in this type of dictionary, but a purely descriptive one.

ICSO, as a *descriptive* dictionary, does not apply criteria that are very selective. According to ICSO rules, one main selection criterion is applied, that is, new words are those units that are documented in use, on the Internet, and that are

[1] "Sin embargo, a pesar de esta oposición conceptual, las relaciones entre la actividad lexicográfica de carácter descriptivo y la de carácter normativo son bastante complejas; incluso se ha considerado que los límites reales entre un tipo de diccionario y el otro no son del todo precisos." (Rafel 2007: 20).

[2] This type of information is also found in other descriptive dictionaries, such as DDLC (Rafel & Soler 2016).

not listed in the lexicographical corpus of exclusion constituted by 7 general dictionaries: DCR, DEX, DEXI, DN, MDA, MDN and NODEX. Note that all these dictionaries are in electronic form and allow a relatively easy search. However, in general, words that do not have at least 10 occurrences in different sources on the Internet are excluded. This condition has been established because on the Web there are numerous sites obtained through automatic translations, which contain words that do not actually belong to the Romanian language or to the Romanian natives. No other selection criteria are applied. Thus, we also introduce words with analyzable structure, which may seem trivial, because this is the Romanian lexicographic tradition and because it is useful for NLP tools. In addition, as Langemets et al. (2019: 9) note, the fact of defining and exemplifying such words is useful for L2 learners. It is worth mentioning that we introduce in the dictionary even the new words and meanings that have been criticized in language cultivation shows, because they are frequent.

3 The Building Procedure and the Entry Structure

Because our institute does not have an IT department that facilitates the (semi-)automatic search of new lexemes and due to the lack of fundamental electronic language resources (exclusion lists, large corpora, reliable IT tools, etc.) the new lexemes / meanings selection is done manually, by the lexicographers involved in the project. This method differs from (semi-)automatic methods (Klosa &Üngen 2018, Kerremans et al. 2012, Falk et al. 2014) in that it does not restrict the search to previously fixed sources to which regular crawling is applied, but it addresses any source on the Internet that uses Romanian and which, preferably, does not represent (automatic) translations from other languages. Online sources can be mass media, blogs, company sites, stores, forums, portals, books, social media etc. In this project, we select single words, multi-word expressions, abbreviations, relevant proper names that serve as derivation bases (e.g. Facebook, Instagram, Nobel, etc.), new elements of word formation (e.g. e- "electronic", robo- "robotic", etc.) and, also, new meanings of older words, provided they gather at least 10 occurrences from different sources and are not found in the exclusion corpus. Each of these constitutes a separate entry in our dictionary, regardless of the existence of morphological or semantic links between them.

Even if the selection is done manually, it is not done in the traditional way, as described by Kilgarriff et al. (2015: 196), that is, by 'reading and marking': "Lexicographers read texts which are likely to contain neologisms – newspapers, magazines, recent novels – and mark up candidate new words, or new terms, or new meanings of existing words". About this method the cited authors claim that it is a low-recall approach, because lexicographers cannot read everything, so there are many neologisms that will be missed. Unlike this method, with low-recall, we instead use a random selection method. Each lexicographer writes down every word seems to be a candidate new word (or meaning), regardless of the communication context in which it is discovered (Web, television, individual talks, personal lectures etc.). If the candidate word is not in the exclusion corpus, we look for further attestation on the Internet, where we find other contexts – of which we choose examples for ICSO – which, in turn, may contain other new words. This way, one can get clusters of new words related by a certain domain or social layer or topic etc. For instance, from a Romanian news story about setting up a skateboard park in a certain locality, a cluster of elements specific to this sport was obtained, e.g. ollie (box), trick, freestyle (board), (grinding) rail, bank, quater pipe, etc. It is worth mentioning that in the list randomly built by lexicographers only about 10% of the candidate words already existed in the exclusion corpus. This gives a hint about the high-precision of the lexicographers' selection, favoured also by the fact that the Romanian explanatory dictionaries do not have, until ICSO project, a consistent and sustained updating program. Furthermore, the lexicographer is not required to read large amounts of text from predetermined sources, hoping to find new words, but they are detected in everyday life according to personal interests. Metaphorically speaking, new lexemes come to the lexicographer provided that he/she pays constant attention to them. Finally, each member of the project team has his/her own list of candidates. Team members' lists may partially overlap, but so far, the number of overlapping words has been relatively small (around 10-15 words in multiple lists). One criticism that could be made of this method is that it does not involve a systematic search and that the precision and recall of the method cannot be accurately calculated. This is true, but we believe that this method is the fastest, least demanding and most productive in terms of the number of new lexemes detected from a source as wide as the Internet. This method could be compared to others, based on the number of new lexemes inventoried in a year with an equivalent workforce, but we do not know such evaluations of other methods.

The inventory is built using the Professional Lexicography Software TshwaneLex (tshwanedje.com). This lexicographic editor has many facilities but we only mention the XML editor, the possibility of RTF or HTML export and the so-called WYSIWYG ("what you see is what you get") view. These facilities help to easily get the dictionary in machine-readable and printed formats. We build the inventory in XML format and in the form of an ODBC database, remotely accessed by the lexicographers (see Figure 1). This format helps lexicographers to see, any time, the whole work and to better collaborate. Moreover, the risk that more lexicographers select and work the same entries independently is eliminated.

As can be seen in Figure 1, the workspace in TshwaneLex is divided into 3 main areas. The area on the left contains the list of all (completed or not) entries in the database. The middle area has a part at the top where the XML elements suitable for the written entry are chosen, according to the hierarchical structure defined in the DTD (Document Type Definition). The selected XML elements are filled with content at the bottom of the middle area. The entries in printed form are displayed in the left area. Entries preceded by a small padlock show that they are "locked" in the database and cannot be opened as long as another user is working on them, so as not to create conflicts. It should be noted that lexicographers do not have exclusivity in writing an entry, the editor or other colleagues may intervene in entries made by others. Of course, the database allows access to remote lexicographers, being very suitable for homework specific to this historical period and also allows access to several users at the same time.

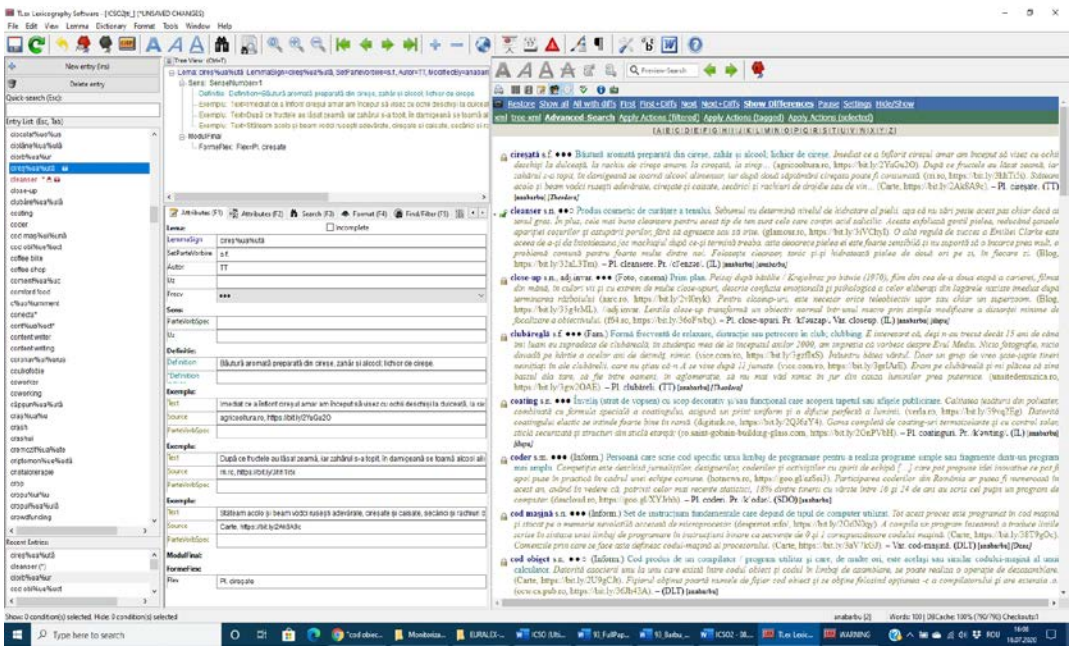

Figure 1: TshwaneLex Framework.

ICSO has the usual entry structure of a general dictionary, see Figure 2. On the lemma, the stressed vowel is indicated, if the lexeme is pronounced as in Romanian, otherwise the pronunciation is indicated in a special field. After the lemma, there is a series of parts of speech associated with it. If necessary, the part of speech is resumed in each sense. For the parts of speech obtained by conversion, no separate senses are described, but only specific examples are given, preceded by the respective part of speech (see *adv.* in Figure 2). The following field indicates the frequency of the lemma on the Internet. This is done with a very rough approximation, because the real dimension of the Internet is completely unknown and an exact count of valid occurrences is impossible. We use 3 symbols: ●○○ (i.e. low frequency) – if the lemma has been found (with Google) on less than four Internet pages; ●●○ (i.e. medium frequency) – if the number of pages containing the lemma is placed between four and ten pages, and ●●● (i.e. high frequency) – if the lemma is present on more than 10 pages. The number of pages is, of course, indicative, because the linguist must eliminate the results in which the lemma appears, for example, in company names, as search tags, in non-Romanian texts, etc. Despite this very rough numerical approximation, we consider that, lexicographically, this frequency information is useful.

The next field, after the frequency, is dedicated to usage information that may refer to selective restrictions, domain, stylistic register, etc. This is followed by the definition of the lemma, expressed by paraphrase and / or synonyms. We try as much as possible not to use, in the definition, words from the same lexical family as the lemma, in order to make the definition as clear and independent as possible. After the definition, a few examples are given that reflect the use of the lemma (and its variants, if any). Each example is accompanied by its online source, consisting of two fields: the web domain and an abbreviated link obtained with the public applications goo.gl or bit.ly. The web domain can provide a good hint about the level of education expected from the language of the text. For example, if this domain belongs to a television station or public institution, one expects the language to be more elevated than if it belongs to a comment on a post or to a forum. The abbreviated link allows the reader to go directly to the site from which the example was taken, to see the expanded context. Another way to get to that site is to search Google for the exact text of the example. It should be noted that choosing the examples is perhaps the most laborious task. This choice must take into account several criteria such as: a) the most credible source of the example (newspapers published in Romania are preferred and sources with Romanian translations are avoided);[3] b) the most appropriate illustration of the meaning (possibly with its explanation); c) the length and clarity of the example. Article titles are also accepted as examples and it should be noted that sometimes a new lexeme appears only in the title (probably due to its brevity), but not in the article.

ICSO entries end with a module containing 4 properly marked fields:

1. Immediately after the symbol "–" the standard inflectional forms are given (if any), for nouns: the plural, and for adjectives: feminine singular, masculine and feminine plural. If one of these is not actually used, the symbol ^ precedes the unattested form. For verbs, the present indicative form of the first or third person and the participle form are specified.
2. After the mark "Pr.", the pronunciation of foreign words or an accent variant (see Figure 2) is given. IPA symbols are used for pronunciation, but only those specific to the Romanian language, because we assume that an ordinary Romanian speaker is not sensitive and would not know how to pronounce the sounds he is not used to.
3. After the mark "Var.", the circulating variants of the lemma are indicated. Often, these variations reflect loanword adaptation trends, such as reduction of double consonants, e.g. *contactles* in Figure 2. Variants are also listed, with reference to the basic form of the lemma.

[3] The Romanian language is also spoken in the Republic of Moldova, but there are obvious differences between the language varieties spoken in the two states.

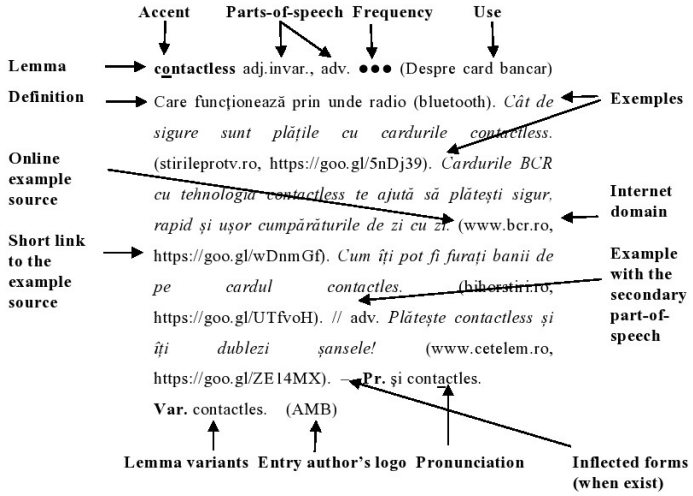

Figure 2: The Entry Structure.

4. After the "Abr." mark, a possible abbreviation of the lemma is given, for example RV for *realitate virtuală* 'virtual reality'. Abbreviations are described as standalone entries that have the same definition as the corresponding abbreviated lexeme and examples focusing on the abbreviation uses.

The last information in the entry refers to its author through an individual logo, e.g. (AMB) in Figure 2. This is an innovation in the lexicographic practice. For communication between lexicographers – both those involved in the project and those who use their results – this innovation has already proved its worth. Other justifications are given below, in section 4.

It should be noted that ICSO does not provide information on the first attestation and on the etymology of the new lexemes. Although, for the very old words, the first attestation is very important, sometimes attracting ample studies to establish it, nowadays, the first attestation is not so relevant. In fact, it is very difficult to establish it in such a large volume of texts. An approximate dating of a new lexeme may be made, in the future, depending on its appearance in a dictionary such as ICSO. A regular and sustained series of a dictionary of new words can be a good benchmark for the first attestation of a lexeme or, more precisely, for the moment of its entry into the language. As for etymology, we have omitted this information so as not to slow down the work. Establishing an etymology is not as trivial as it seems and requires specialized research that is time consuming.

The lemmas that already exist in the exclusion corpus are marked with the symbol *. They are entered in our dictionary if they have new meanings or the definitions in the other dictionaries are (no longer) appropriate or if they present any other element of novelty (e.g. changes in spelling, selective restrictions, etc.).

4 First Results and Encountered Problems

During two years (including one year for the project design and lexicographic rules setting), about 1400 entries have been built by 5 lexicographers working part time (1,5 full time).

Regarding the lexical creativity reflected by the new lexemes, we do not intend to make a detailed description here, but only mention a few aspects. It is worth mentioning that most of the inventoried lexemes, at this stage, represent lexical creations and extensions of lexical families in Romanian, and the loanwords in original form and their naturalized derivatives cover only 20% (despite the general perception of the overwhelmingly invasion of anglicisms, for instance).

Another aspect concerns a whole series of words that, most likely, will not last in time, but which reflect a notable social phenomenon: the strong influence of highly publicized people on society. Lexical families of some proper names belonging to persons very frequently seen on TV were registered. For example, the name of a controversial businessman, Becali, created the family made of adj. *becal-ic* / *becal-ist* 'specific / loyal to Becali', vb. *becal-iza* (infinitive) / *becal-izat* (past participle) 'become as Becali'. Of course, the person's name Becali is not entered as a separate entry, as we consider that the reference in the definition to the person with that name is sufficient.

Graphic adaptation of loanwords can be traced with the help of the registered variants of different lexemes. They reflect the following types of adaptations, among others:

- phonetic writing, specific to the Romanian language: *brandui* 'to brand' > *brendui*, which is written as it is pronounced; *flash* > *fleş*; *foosball* > *fusbal*; *rider* > *raider*; *vlogger* > *vlogăr*; etc.
- deleting letters that are not pronounced (double consonants, mute vocals etc.): *contactless**s*** > *contactles*; ***co**uponing* > *cuponing*; *fratel**l**o* > *fratelo*; *pat**t**ern* > *patern*; etc.
- writing in ordinary Romanian letters: ***cy**berterorism* > ***ci**berterorism*; *flas**h*** > *flaş*; ***ph**oto**sh**opare* 'photoshopping' > ***fotoşopare***; etc.
- changing the English plural ending to the Romanian one: *dreadlock**s*** > *dreadlock-**uri***, etc.

In fact, the vast majority of variants reflect hesitations in using the hyphen to mark compounds or derived words. For instance, for compounds, especially loanwords, there are almost complete series of variants: *flash-mob / flash mob /*

flashmob / fleșmob; *microjob / micro-job / micro job*; *off-grid / off grid / offgrid*; etc. But there are also hesitations regarding Romanian compounds: *cardiotoracic / cardio-toracic* 'cardiothoracic'; *colorectal / colo-rectal* 'colorectal'; *euroentuziast / euro-entuziast* 'one who has full confidence in the values of the European Union.'; etc. The same phenomenon is found in words derived in Romanian with prefixes: *anti-avort* / *antiavort* 'anti-abortion'; *co-inventator* / *coinventator* 'invention partner'; *interreligios* / *inter-religios* 'interreligious'; or suffixes: *rohmerian* / *rohmer-ian* 'in Rohmer's style'.
The graphic variation can also come from other sources, such as writing abbreviations as they are pronounced: *CAP-ist* / *ceapist* 'worker in an agricultural cooperative'; *PSD-ist / pesedist* 'member in the Social-Democrat Party'; or lowercase / uppercase writing: *new-age-ist* / *New-Age-ist* 'adept of New Age philosophy', *secret Santa / Secret Santa*, *youtuber / YouTuber* 'person who regularly makes and posts videos on youtube.com'.
The recording of all these variants, useful for a deeper understanding of language trends, also creates some problems. The main problem we encountered in building ICSO is related to the fact that our institute is seen as the main author of normative academic workings. Thus, we faced the same problem reported by Joaquim Rafel (see also Langemets et al. 2019: 12):

> Uno de los problemas que plantea la elaboración de un diccionario descriptivo por una academia de la lengua es que este diccionario contiene palabras o acepciones que no son reconocidas por la normativa vigente, a pesar de encontrarse documentadas en los textos; por una parte este diccionario puede ser considerado más científico que el normativo por cuanto intenta dar cuenta de una manera sistemática de la realidad de la lengua a partir de datos empíricos, pero por otra parte puede ser visto como un peligro para el uso lingüístico considerado correcto. (Rafel 2007: 21)

Actually, in general, the public demands linguistic norms even for things that cannot be normed. It is so eager for normative works. Therefore, given this expectation, a descriptive dictionary can create confusion, because it contains many variants under which a lexeme circulates and many terms belonging to the „colourful" language used in forums, comments, etc. To draw attention to the fact that not all forms belong to the literary language, we have adopted the following solutions (in addition to the warning note in the generally ignored introduction).
Firstly, we have called this work "inventory", not "dictionary". The Romanian term "inventory", mainly used in dialectology, suggests a simple enumeration and differs from the usual titles of academic dictionaries.
Secondly, we have paid more attention to the stylistic-use information, such as depreciative, ironic, affectionate, etc. We could not mark the unrecommended entries, as it has been done in Rafel (2007: 21) for instance, because ICSO entries have not been yet subject to prescriptive analysis.
Finally, we have adopted an innovative solution by ending each ICSO entry with its author's logo, in order to highlight the fact that the entry content is under the responsibility of a person, not an authority. For the general public, the personalization of the entry could diminish the normative perception, and for the lexicographer colleagues it helps to a better communication later. The author logo also concerns the loanwords pronunciation which, in the absence of specialized studies, reflects the choice of the entry author.
Another encountered issue is related to the volatility of the examples on the Internet. This is indeed a risk, mitigated by the fact that if a lexeme is certified at least 10 times, it is likely to remain attested even if the example in ICSO disappears. In addition, when selecting examples, archival sources are preferred and volatile texts, like those found on sites of second-hand sales, are avoided.

5 Conclusions and Further Work

ICSO is a descriptive dictionary that monitors new lexemes in Romanian, in order to capture the dynamics of the language in the smallest details (e.g. the loan adaptation process). The main purpose is to provide primary material for the systematic updating of normative dictionaries. A secondary goal is to provide the general public with explanatory definitions for the explosion of terms on the Internet.
In order to draw the public's attention to the fact that this is not a normative dictionary, we have adopted small lexicographical innovations such as the title "Inventory" and the indication of the entry author through a logo. In order to ensure an increased working speed, required by the urgent need to update general dictionaries, we have adopted the random selection method which excludes the regular browsing of predetermined sources. We also waived the information regarding the first attestation and the etymology of the registered lexemes. The use of the Internet as a lexical source allows us access to a field almost ignored by normative dictionaries, that of familiar language in social media, which actually reflects everyday language.
The project aims at the sustained elaboration of a volume of at least 1500 words every 2 years, taking into account the inherent delays due to publication. The publication of the dictionary on the site of our institute is considered, possibly in a dynamic way, so that the public can benefit as soon as possible from the definitions of the new lexemes. Another aspect worth considering is the use of the crowdsourcing method for word gathering. With the launch of ICSO2 site, a facility can be created to allow the general public to suggest new words to be introduced in the dictionary. Furthermore, we hope that in the near future we can use tools for automatic search of new lexemes that have sufficiently good results.

6 References

Adelstein, A. & Freixa, J. (2013). Criterios para la actualización lexicográfica a partir de datos de observatorios de neología. Unpublished presentation at Congreso Internacional El Diccionario: neología, lenguaje de especialidad, computación, Ciudad de México (Mexico), 28-30th October 2013. Accessed at: https://repositori.upf.edu/handle/10230/34891 [04/10/2020].

Barbu, A. M., Croitor, B., Niculescu-Gorpin, A. G., Radu, I. C. & Vasileanu, M. (2020). *Inventar de cuvinte și sensuri noi atestate în mediul online (ICSO 1),* vol. 1. Editura Academiei.

Cabré, M. T. (1993). *La terminología. Teoría, metodología, aplicaciones*. Barcelona: Editorial Antártida/Empúries.

Crystal, D. (2001). *Language and the Internet*. Cambridge: Cambridge University Press.

DCR – Dimitrescu, F., Ciolan, Al. & Lupu, C. (2013). *Dicționar de cuvinte recente*. Editura Logos.

DEX – *Dicționarul explicativ al limbii române* (1998-2016). Institutul de Lingvistică „Iorgu Iordan – Al. Rosetti": Editura Univers Enciclopedic Gold.

DEXI – Dima E. (ed.) (2007). *Dicționar explicativ ilustrat al limbii române*. Editurile ART și GUNIVAS.

Diaconescu, S. Ș. (ed.) (2015). *Fonetica limbii române*, 4 vol. SOFTWIN: CreateSpace Independent Publishing Platform.

DN – Marcu, F. & Maneca, C. (1986). *Dicționar de neologisme*. Editura Academiei.

Falk, I., Bernhard, D. & Gérard, C. (2014). From Non-Word to New Word: Automatically Identifying Neologisms in French Newspapers. In *Proceedings of the 9th Edition of the Language Resources and Evaluation Conference*, May 2014, Reykjavik, Iceland, pp. 4337–4344.

Freixa, J. & Torner, S. (2019). Beyond Frequency: On the Dictionarization of New Words in Spanish. In *Kernerman Dictionary News* 27, July 2019, p. 6. (Presentation at the Globalex Workshop on Lexicography and Neologism, Bloomington, Indiana, US, 8th May 2019.) Accessed at: www.academia.edu/39070136/ [04/10/2020].

Fuertes-Olivera, P. A. (2012). Lexicography and the Internet as a (Re-)source. In *Lexicographica* 28(1), pp. 49-70.

Fujii, A. & Ishikawa, T. (2000). Utilizing the world wide web as an encyclopedia: Extracting term descriptions from semi-structured text. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, *ACL-2000, 3-6 October 2000,* Hong Kong, pp. 488-495.

Gouws, R. H. & Prinsloo, D. J. (2005). *Principles and Practice of South African Lexicography*. SUN Press.

Grefenstette, G. (2002). The WWW as a Resource for Lexicography. In M.-H. Corréard (ed.) *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*. Grenoble, France: EURALEX, pp. 199-215.

Kerremans, D., Stegmayr, S. & Schmid, H.-J. (2012). The NeoCrawler: Identifying and Retrieving Neologisms from the Internet and Monitoring Ongoing Change. In K. Allan, & J. A. Robinson, (eds.) *Current Methods in Historical Semantics*. De Gruyter Mouton, pp. 73-59.

Kilgarriff, A. (2001). Web as corpus. In *Proceedings of Corpus Linguistics 2001*. Lancaster, UK(March). Reprinted in G. Sampson & D. McCarthy (eds.), *Corpus Linguistics. Readings in a Widening Discipline*. 2004. London and New York: Continuum, pp. 471–473.

Kilgarriff, A., Herman, O., Bušta, J., Kovář, V. & Jakubíček, M. (2015). DIACRAN: a framework for diachronic analysis. In *Corpus Linguistics 2015*. *Abstract Book.* Lancaster: UCREL, pp. 195-197.

Klosa, A. & Lüngen, H. (2018). New German Words: Detection and Description. In J. Čibej, et al. (eds.) *Proceedings of the XVIII EURALEX International Congress*. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 559-569.

Langemets, M., Kallas, J., Norak, K. & Hein, I. (2019). New Estonian Words and Senses: Detection and Description. In *Kernerman Dictionary News* 27, July 2019, p. 8. Accesed at: globalex.link/wp-content/uploads/2019/05/gwln2019_langemets-kallas-norak-hein.pdf [04/10/2020]

MDA – *Micul Dicționar Academic* (2010). Institutul de Lingvistică „Iorgu Iordan – Al. Rosetti": Editura Univers Enciclopedic.

MDN – Marcu F. (2000). *Marele Dicționar de Neologisme*. Editura Saeculum.

Naparsteck, M. (2005). *Honesty in the Use of Words*. Rochester, New York: Lake Affect Publishers.

NODEX – *Noul Dicționar Explicativ al Limbii Române* (2002). Litera Internațional: Editura Litera Internațional.

O'Donovan, R. & O'Neill, M. (2008). A systematic approach to the selection of neologisms for inclusion in a large monolingual dictionary. In E. Bernal, J. DeCesaris (eds.), *Proceedings of the XIII EURALEX International Congress*, Barcelona, 15-19 July 2008. Barcelona: IULA-UPF, pp. 571-579.

Rafel i Fontanals, J. & Soler i Bou, J. (2016). A Descriptive Dictionary of Contemporary Catalan: The DDLC Project. In E. Corino, C. Marello, C. Onesti (eds.) *Proceedings of the XII EURALEX International Congress*, vol. I, Turin, Italy, 6–9 September 2006. pp. 443-455.

Rafel i Fontanals, J. (2007). Prescripción y descripción en la actividad académica: el Diccionari descriptiu de la llengua catalana. In M. Campos et al. (eds.) *Reflexiones sobre el diccionario, Actas del I Congreso Internacional de Lexicografía Hispánica*, Coruña: Setembre 2004. Universidade da Coruña, pp. 9-33.

Sánchez Manzanares, C. (2013). Valor neológico y criterios lexicográficos para la sanción y censura de neologismos en el diccionario general. In *Sintagma* 25, pp. 111-125.

Zgusta, L. (1971). *Manual of Lexicography*. (Janua Linguarum Series Maior 39). Prague/The Hague: Academia/Mouton.