Journal of Portuguese Linguistics

Fokin, Serhii. (2025). Semiautomatic selection of interjectional onomatopoeia from English, Portuguese, Spanish, and Ukrainian corpora based upon syllables' repetition pattern. *Journal of Portuguese Linguistics*, 24, pp. 1–27. DOI: https://doi.org/10.16995/jpl.15242

Open Library of Humanities

Semiautomatic selection of interjectional onomatopoeia from English, Portuguese, Spanish, and Ukrainian corpora based upon syllables' repetition pattern

Serhii Fokin, Taras Shevchenko National University of Kyiv, UA, sergiyborysovych@ukr.net

Onomatopoeic words constitute a serious challenge for translators, lexicographers, language learners, and teachers. Hence, empirical data collection on onomatopoeia is highly sought after. The most suitable data sources for extracting onomatopoeia are large language corpora. Since onomatopoeic words and, particularly, interjectional onomatopoeias show wide variance and many of them are created spontaneously, the methodology chosen for automating the extraction in this study initially involved observing the existing patterns of transcribed interjectional onomatopoeias, among which the one based upon repetition proved the most recurring. Among the observed features were the same or similar syllable sequence, three or more repeated letters, combined with punctuational markers such as hyphens, ellipses, quotation, or exclamation marks, part of speech tags. These properties were further implemented in formulating corpus queries. The search was based on the pattern of repetition of similar syllables. The results underwent an ANOVA test that revealed the open and closed hyphenated syllables to be the most reliable pattern for extracting interjectional onomatopoeias from corpora of English, Portuguese, Spanish, and Ukrainian. The used markers allow for the achievement of high efficiency, which was evaluated in terms of precision.

1. Introduction

Phonetic motivation as a lexeme creation mechanism constitutes a considerable theoretical and practical challenge. Onomatopoeic words that may appear spontaneously in a given language and are currently widely used are characterized by a high degree of occasionality. Their examples in written literature are rare, whereas their usage is not clearly regulated, excepting the most frequent forms traditionally mentioned in dictionaries and grammars: *bow-wow*, *bang*, *tic-tac*, and similar. One of the practical challenges that arises from this extravagant phenomenon is their correct usage in foreign language learners' speech: for whatever correct grammar and vocabulary is used in spontaneous speech, inappropriate onomatopoeic words are likely to unveil the speech's unnaturalness. In the domain of translation practice, the onomatopoeia usage and meaning are a poorly explored subject, as this phenomenon posits a practical challenge. As Casas-Tost points out:

> As I see it, one of the factors which are an encumbrance to the translator's task is that these text units have been given little importance at a theoretical level and, as a consequence, in practice. This is reflected by the lack of onomatopoeia entries in all manner of reference books, including dictionaries, which I believe is one of the reasons why they are rarely used (2012, p. 39).

It is evident that many onomatopoeias are used *ad hoc* and are highly dependent upon the situational context. There are hundreds of conventional onomatopoeic words used in fiction, in transcribed oral texts, and in internet communication that could be additionally registered in monolingual and bilingual dictionaries and that would be of a high practical value for language learners, teachers, and translators. The specialized literature is characterized by scarce observations in this respect, since examples are hard to find.

However, it should come as no surprise that it is possible to find mentions of lexicographic sources focused particularly upon onomatopoeia in either monolingual or even bilingual dictionaries of a limited set of languages, for example, *Farhange Namavaha dar Zbane* ("A dictionary of Onomatopoeia in Persian") by Vahidian Kamyar (1996). Curiously, bilingual or multilingual dictionaries that concern this subject are more prolific than monolingual ones—perhaps because the dictionary compilers became aware of the object's importance through specific translation or language-learning challenges. Such works include *Diccionari d'onomatopeies i altres interjeccions: amb equivalències en anglès, espanyol i francés* by Riera-Eures (2010) for Catalan, English, Spanish, and French, and *Japanese-Ukrainian Themed Dictionary of Onomatopoeic Vocabulary* by Egava & Kobelyanska (Егава, 2016), which offers the user a wide range of search possibilities from alphabetical criterion to accessing through subject classification (being this onomasiological approach still quite rare among lexicographers). After a detailed overview, Medvediv and Dmytruk (2019, p. 79–80) provide an extensive list of the Japanese lexicographers' achievements. Despite these examples, for most languages and language pairs, the lexicographic gap of onomatopoeia is still not covered.

In spite of the emerging literature and lexicographic sources regarding onomatopoeia, this lexically, emotionally, culturally, communicatively, and stylistically remarkable feature

still constitutes an impressive lacuna in the domain of language didactics, lexicography, and translation due to data shortage.

The logical question arising in similar cases is whether there is a possibility of automating the selection in large-volume data, such as language corpora. Therefore, the purpose of this study is to find a method of automating the extraction of onomatopoeic words by observable formal markers, and evaluate its effectiveness in terms of precision. To be able to generalize commonalities in such markers, we resort to annotated corpora of four different languages: English, Portuguese, Spanish, and Ukrainian.

The article is organized as follows. After this introduction, in Section 2, Theoretical background, we explore the approaches to the subject in literature regarding onomatopoeia as a translational and lexicographic challenge, as well as methods of automatic retrieval of onomatopoeia. In Section 3, Methodology, we elaborate upon the methods used to automate the onomatopoeia extraction from corpora of English, Portuguese, Spanish, and Ukrainian. In Section 4, Results and Discussion, we propose tools to evaluate the precision of the performed corpus queries and judge the statistical significance of the expected precision rates obtained in the course of the study. Finally, in Section 5, we summarize the most common formal properties that may be successfully used in corpus queries to extract interjectional onomatopoeias.

2. Theoretical background

2.1. Interjection VS onomatopoeia: distinguishing criteria

"*Onomatopoeia* is the naming of a thing or action by a vocal imitation of the sound associated with it (such as *buzz* or *hiss*)" (Britannica, 2024). The phenomenon in question is not restricted to a particular part-of-speech (POS), however due to lexical and functional similarities researchers tend to associate onomatopoeia primarily with interjections, which is seen in some manuscript titles, such as *Diccionari d'onomatopeies i altres interjeccions: amb equivalències en anglès, espanyol i francés* by Riera-Eures (2010). In fact, sophisticated criteria are needed to distinguish both terms from each other, which is why entire works, even PhD theses, are devoted to this issue (Meinard, 2022).

Interjection, on its turn, is defined as "an exclamatory word or phrase used to express an emotional reaction or to emphasize a thought" (Britannica, 2024). Once compared the definitions of these narrowly interrelated terms, we can conclude that the differences between both phenomena lie in their semantic meaning: while the onomatopoeia expresses sounds, the interjections convey emotions. Rodríguez Guzmán infers a set of additional points of inflection (formal characteristics, syntactic function in sentences) and disjunction (motivation patterns, morphonological processes, semantics) (2011, p. 173) between onomatopoeia and interjection, and finally concluding that both are to be considered as separate word classes (2011, p. 173). If we interpret the *word class* as *part of speech* in the context of data mining, particularly, in corpus linguistics, most corpora managers and taggers rely on worldwide conventions, among which the most widespread is the *Universal Dependencies* (UD) framework, currently used to annotate

thousands of corpora. While the UD POS tagset does include interjections, the onomatopoeia does not form its part (Universal Dependencies, 2024). Similarly, many other traditional POS lists, whose number typically ranges between 9 and 10, comprise interjections (since the Latin grammars) but traditionally exclude onomatopoeia. This seems consistent with the logic: while, indeed, onomatopoeia semantically stands alone from other POSs, expressing sounds, the semantic category of a word is not the decisive criterion to assign it a POS label: otherwise, verbal nouns, such as *participation* or *engineering* should be semantically classified as verbs.

Both grammar and semantic characteristics come into play when assigning a POS property to a word. From a grammatical standpoint, the onomatopoeias are unchangeable words, as well as the interjections. Furthermore, depending on a particular linguistic school, the sound-imitating meaning is listed among the semantic properties of interjections, which is the traditional approach in Ukrainian grammar (see, for instance, Karamysheva, 2017, p. 218). This feature is also empirically validated by the sampling from the corpus GRAK (see **Table 9** and **Table 10**), where including the interjection tag in the query yields an impressive number of onomatopoeias. In this case, the concept of interjection turns out to be broader than that of the onomatopoeia. Now, the central question is what POS status should apply for *onomatopoeia*? Beyond the ongoing debates about their part of speech status, there is an immediate need to retrieve sound-imitating words from corpora or another source for various practical purposes. Authors, translators, and editors may need to express sound not only using pure sound imitation but also through derived nouns, verbs, adjectives, and adverbs with the semantics of sounds. Are these words to be classified as onomatopoeias? According to the *Merriam-Webster Dictionary*, onomatopoeia can also refer to the words formed by onomatopoeia (2024). Thus, not only *buzz*, *hiss* and similar words, but also *buzzing*, *hissing*, *buzzy*, *and hissy*, as well as other onomatopoeically derived lexemes (nouns, verbs, adjectives, adverbs), may form part of this list, as seen (particularly, but not exclusively) in Bidaud's work (2022), who focuses their research on *verbal onomatopoeias*. Moreover, in English, assigning a part of speech property for a word such as *buzz* may be particularly challenging. It is obvious now that the POS-attribution may depend on the semantic and grammar approach elaborated upon in a particular linguistic school, but, from the standpoint of data mining at the current stage, we are to conclude the following:

1) while the interjection is universally considered a part of speech, onomatopoeia is not;
2) onomatopoeia is now qualified rather as a semantic word class that may be assigned different part of speech tags;
3) interjection is a class that may comprise onomatopoeia depending on the linguistic approach;
4) both classes (onomatopoeia, interjection) in their broadest sense intersect; in the intersection of both classes, emerges a subclass of interjectional onomatopoeias.

Given these premises, we qualify hereafter *interjectional onomatopoeias* as onomatopoeias morphologically characterized as interjections, which act as grammatically unchangeable words and semantically express sounds.

2.2. Onomatopoeia as a translation challenge

It is important to note that the issue of translating onomatopoeic words comprises two faces: whereas in translation many context-driven techniques allow for multiple ad hoc contextual solutions, for the sake of dictionary compilation, more comprehensive solutions covering a wide range of potential situations in translation are needed.

From the translatological point of view, Yaqubi et al. outline a dozen techniques for using sources to translate onomatopoeia in the following order of precedence:

> Make every effort to apply "established translation", in other words, choose the exact recognized equivalent in the dictionary. In case, no equivalent is chosen for the item, choose "discursive creation" technique in order to create the same effect, although out of the context of the literary work, it may not have the same effect. Use "borrowing" technique which can help the translators to transfer the expressive function of the onomatopoeias to some extent. This transference is due to the universality of sound effects. Use "descriptive translation" in TT in order to imply that the item imitates a sound and the sound implies an action or emotion. Utilize "generalization" which helps to transfer the general meaning of onomatopoeias, i.e., the general information about the action and the emotion by using specific lexicon. However, by using this technique the form used in TT may not sound like onomatopoeia. Apply "reduction", although by using it, some information may be partially or completely missed by the translator. Mix the translation techniques in order to create an equivalent which can imply the expressive function both in form and meaning (2018, p. 220).

The mentioned procedures are of enormous practical help for translators who are constrained to bridge numerous onomatopoeic lacunas. Whereas in one language there may be a traditional way of imitating the sound of a particular object or phenomenon, in other languages there might not exist a conventionally accepted onomatopoeia for a given communicative situation, where in yet another, less specific onomatopoeia might be acceptable (i.e., by means of hyperonymic substitution). The onomatopoeia of scissors or other cutting tools in Ukrainian is *чик-чик*, although no specific sound-imitating word is provided in the list of onomatopoeic words for Spanish, whereas in English and Portuguese there seem to exist analogous onomatopoeias: *snip*, and *rip*, as shown in the following examples 1 and 2.

(1) These ninjas with scissors often have the vision to see the best version of you long before you can see it yourself. And there is nothing quite like the anticipation of patiently sitting in a chair, hearing the ***snip-snip*** of the scissors and watching a new you emerge in the mirror (https://www.rsvplive.ie/life/hairdressers-unsung-heroes-lives-writes-14097868) (RSVPLive 2024).

(2) Então ela agarrou os lindos cabelos de Rapunzel, deu-lhe algumas palmadas com a mão esquerda e com a direita apanhou a tesoura e ***rip, rip, rip***, os cabelos estavam cortados (Chamizo Babo n.d., *CRPC*).

At the same time, a similar sound in Spanish can be traditionally imitated through *zas-zas*, whose meaning denotes many types of noise, i.e., is a hyperonym (see example 3):

(3) ***¡Zis, zas y zas!*** Una y otra vez zarandeó tijereteando el gladio vorpal! Bien muerto dejó al monstruo, y con su testa ¡volvióse triunfante galompando! (A bordo del Otto Neurath n.d.).

Beyond any possible valuable technique (reproduction, substitution, addition) that may serve as a brilliant situational workaround, it is crucial to explore the existing bilingual and monolingual dictionaries of onomatopoeia first to discover the possible lexicographic gaps and attempt to bridge them, very much in accordance with Yaqubi's et al. recommendation of looking for an established equivalent as a priority method (2018, p. 220). Nevertheless, borrowing onomatopoeias from the source into the target text looks like a widespread technique. Rodríguez Guzmán presents a list of onomatopoeic words loaned and borrowed into Spanish from other languages (2011, p. 157). At the same time, some tricks used in translation due to the absence of a better suitable equivalence in the target language may indeed be due to a gap or lack of knowledge of existing specific means. This implies that specialized informational resources (such as dictionaries or corpora) would be of great help for translators.

2.3. Onomatopoeia in bilingual lexicography

It is obvious that, to compile bilingual or multilingual dictionary entries, credible sources and robust methodology are required. From the standpoint of the 21th century lexicography, the undisputed number one source for extracting extensive linguistic data are large-volume language corpora, and onomatopoeia is not an exception. In fact, many researchers use their own custom corpora to perform manual searches. For instance, Yaqubi et al. use their research corpus to calculate the frequency of the onomatopoeia in the Charles Dickens novel *A Tale of Two Cities* and for performing manual searches to retrieve examples of onomatopoeia for their two translations into Persian (2018, p. 211–212).

Some papers do pursue the objective of automating this operation: Orrequia-Barea and Marín-Honor explore techniques particularly focused on onomatopoeic word extraction from large-volume corpora, such as the British National Corpus (2020, p. 47). Nevertheless, works of this kind are few, and our objective is to propose a method to optimize retrieving onomatopoeic words from large corpora and to evaluate their effectiveness in terms of precision involving four languages: English, Spanish, Portuguese, and Ukrainian.

3. Methodology

3.1. Observation

To find rational ways of retrieving examples of onomatopoeia, first we need to find out their most relevant features. As a starting point for observation, we have primarily used the ready-made lists of onomatopoeias in English (Yourdictionary, 2021), Portuguese (Riondlearn, 2022),

Spanish (Fundeu, 2011), and Ukrainian (Божко, 2023). These lists were subject to observation with the purpose of extracting some valid markers to be used as reliable formal criteria during automatic or semiautomatic extraction of onomatopoeic words out of corpora. The method of observation, aimed at an intuitive selection of relevant features, yielded the following recurrent (but not exclusive nor mandatory) characteristics of the onomatopoeic words:

1) repeated mostly closed syllables with the same vowels (*pam-pam*);
2) repeated open syllables (*chu-chu*) and repeated syllables with different vowels (*zigzag, flip-flops*);
3) observed repetitions are mostly hyphenated, but merged forms are not rare (*tacatar, toc-toc-toc, ronroneo, tantan, pompom*);
4) repetition of three or more graphemes (*zzzzz, piiuuw, zwiiiz, pionggg*);
5) ending of the word with *-h* (*pouah, schh, pchhh*).

The second and fourth observations partially coincide with the patterns proposed by Orrequia-Barea and Marín-Honor (2020, p. 52). The repetition in linguistic sounds' representation or, specifically, reduplication are known as a universal linguistic feature:

> The repetition of sounds occurs in all languages of the world, doubling segments of audible material: natural sounds and animal cries, but also words and clauses (…) it is interesting to note how often reduplication serves as a common denominator even in cases when languages disagree in the choice of phonemes (Anderson Earl 1998, p. 112).

The fifth observation, once implemented in corpus queries, did not produce noteworthy results. Each of the other four observed items merit particular attention and study, and we implemented the detected features in corpus queries. In the current study, our purpose is to focus upon the syllable's reduplication pattern and the possibilities of its usage to automate onomatopoeia extraction from corpora. Therefore, the methodology is based on the phonic properties of onomatopoeias, such as repeated syllables, and, where possible, upon the part of speech parameters. Since there are no phonically annotated large corpora of the languages in question (Ukrainian, English, Spanish, and Portuguese), we were constrained to base our queries on grapheme levels instead of sound or phonemes.

To achieve the results, we use particular query languages depending on the search engine the corpora are provided with the *Corpus Query Language*, *CQL* (Sketch Engine, n.d.) or *Corpus Query Processor, CQP* (Evert S. & The CWB Development Team, 2022) because they allow for searching patterns matching both regular expressions and linguistic annotation tags.

3.2. Former empirical researches

The automation of onomatopoeia retrieval as an idea started being explored in 2020. Although our interest in the subject appeared independently from the existing studies using similar

techniques, we retrospectively took into account all the achievements that are very much in accordance with the current paper.

Orrequia-Barea and Marín-Honor (2020) systematize different graphic properties of onomatopoeic words in written text to extract them by using the regular expression syntax, with some interesting observations regarding the correlation between onomatopoeic formal pattern and the ontological nature of the represented sound using Round & Kwon's concept of *phonaesthemes*, i.e. "recurrent pairings of sound and meaning" (Round & Kwon, 2015, p. 2).

Orrequia-Barea and Marín-Honor searched the text of comics to retrieve onomatopoeia in Spanish and French, and for texts in English the regular expression syntax was applied:

> All the above-mentioned systematisations were captured by means of regular expressions, which are patterns that are frequently used in text editors to look for phonaesthemes. This sequence has to fulfil the criteria set out by the regular expression. As the main purpose was to find most of the onomatopoeias in the BNC, the following regular expressions, based on the previous patterns of formation, were used: 1. To find consonants that were repeated at least three times: [bc-df-hj-np-tvz]{3}. This regular expression yielded onomatopoeias such as *zzz*. 2. To find the pattern of up to two consonants plus vowels, repeated at least twice, followed optionally by an indefinite number of consonants: [bc-df-hj-np-tvz]{0,2}vowel{2,} [bc-df-hj-np-tv-z]{0,}. We typed each of the five different vowel graphemes in the vowel slot. Some of the results were: *craark*, *beep*, *riing*, *boom* or *uuummm* (2020, p. 52).

More observation on potentially universal features of onomatopoeias comprising some sound combinations are described by Assaneo, Nichols & Trevisan:

> We explore the vocal configurations that best reproduce non-speech sounds, like striking blows on a door or the sharp sounds generated by pressing on light switches or computer mouse buttons. From the anatomical point of view, the configurations obtained are readily associated with co-articulated consonants, and we show perceptual evidence that these consonants are positively associated with the original sounds. Moreover, the pairs vowel-consonant that compose these co-articulations correspond to the most stable syllables found in the knock and click onomatopoeias across languages, suggesting a mechanism by which vocal imitation naturally embeds single sounds into more complex speech structures (2011, p. 11).

Some researchers utilized dictionaries as a starting point for retrieving the onomatopoeic words out of lexicographic sources explicitly marked as onomatopoeic:

> I made a list of onomatopoeic words using the following three steps. First, I searched entries (i.e. head words) in the OED, including terms such as onomatopoeia/onomatopoeic/onomatopoetic etc. in their etymologies. Specifically, I typed onomatop* into the "FIND WORD" box in

> the advanced search of the OED and restricted the search area to etymologies. As a result, 385 entries met this condition. (…) However, the list of these 304 entries is not adequate in itself. The OED often treats different grammatical classes of one word (= lemma) as separate entries. In addition, these separate entries are sometimes not given the same explanation of their etymologies. Many entries would be overlooked if I examined only those entries that included onomatopoeia/onomatopoeic/onomatopoetic etc. in their etymologies (Takashi Sugahara, 2011, p. 34)

A similar approach was implemented by Yaqubi et al. (2018, p. 212) who compiled a list of Persian onomatopoeias at the initial stage of their research. Although this methodology seems promising, we did not apply it in the current study, for it implies significant manual work and does not allow for extraction automation. Moreover, at the current stage, where numerous onomatopoeic lemmas or their graphic variants are not yet included in dictionaries, we chose to extract sound-imitating interjection from corpora including occasional ones. This is why we opted for patterns based on syllables or grapheme repetitions.

We are aware that the patterns based upon repetitions ignore single-syllable words. However, the basic idea relies upon the fact that the repetitions may serve as an access point to further retrieve another kind of onomatopoeia that is not based upon repetitions, some of which are present in the nearest context. In other words, if an onomatopoeia exists in the form of repeated syllables, then it is likely to appear in its monosyllabic or isolated variant, as seen in the following examples (4, 5 and 6) in Portuguese and Ukrainian:

(4) Tenho a favor deste meu juízo o facto de que, tendo o Governo calculado tam modestamente o rendimento deste imposto em 10 : 000 contos, ele veio a render 150: 000 – diz a Câmara Corporativa -, mas há por aí uns ***zuns-zuns*** que dizem que chegou a 200: 000 (http://gamma.clul.ul.pt/CQPweb/crpc/textmeta.php?text=A25999&uT=y).

(5) O Ângelo veio-me para cá com, uns zum ***zuns*** (NEMÉSIO, Vitorino, 1944, *CRPC*).

(6) У нас було по 130 поранених удень. Їх не тільки я вивозив, звісно. Але уявіть, що тут робилося. “***Бах***! ***Бах***! ***Баааааах***!” [We had about 130 wounded per day. Of course, I wasn’t the only me to evacuate them. But just imagine what was happening here. ‘Bang! Bang! Baaaang!] (*GRAK* 2023: *Репортер*, 2022).

After elaborating upon the enquiry method, we needed to choose the best fitting corpora for extracting empirical data among the available corpora provided with the possibility of looking patterns of texts matching regular expressions.

3.3. Choosing corpora

To subject to test our hypothesis by means of corpus query patterns that we further propose, we need to decide what corpora will serve as the source of empirical basis, which is why we resort to

corpora of four languages in which we can read and, consequently, carry out contextual analysis: English, Portuguese, Spanish, and Ukrainian. The number of languages used is also constrained by the accessibility of specific query languages in the corpora interfaces, suitable with the proposed patterns. Although English and Ukrainian are genetically distant from Spanish and Portuguese, the conclusions drawn on four languages from different groups will allow for better judging over the validity of the query patterns. An additional reason to choose English, Romance languages, and Ukrainian was the fact that, for the latter, the bilingual lexicographic contributions in the field of onomatopoeic dictionaries is particularly fruitful, as seen in 2.2., whereas English-Ukrainian, Portuguese-Ukrainian, and Spanish-Ukrainian language combinations are not provided with lexicographic sources containing either interjections or onomatopoeias. Moreover, among the four mentioned languages, only the Ukrainian corpus is provided with correctly tagged interjections, which helps bring to light additional properties of the queries performed.

The best intuitive choice of corpora may seem the referential standard since the reference corpora better represent the general properties of a language. At the same time, with regard to the needs of the research, we are also constrained by the corpora technical details. Orrequia-Barea and Marín-Honor stressed on the downloadability of the corpora:

> Our first idea was to extract onomatopoeias from corpora of each language since we wanted to have empirical evidence that those onomatopoeic forms were actually used in the language. For this reason, we intended to download the corpora to look for onomatopoeias using regular expressions to get as many onomatopoeic forms as possible without restricting them to the most common ones. However, we could only follow this procedure with the BNC, since it was the only corpus that could be downloaded. For Spanish and French, the CREA and FRANTEXT corpora were not downloadable, so that we had to follow a different process, namely manually extracting onomatopoeias from comics (2020, p. 51).

To overcome this difficulty, we resorted to corpora provided with the *CQL* (Sketch Engine, n.d.) and *CQL* query language (Evert S. & The CWB Development Team, 2022), whose usage is illustrated in the next section, which allows for the usage of regular expressions; therefore, it was not mandatory to download any corpus.

While the Portuguese reference corpus *CRPC, Corpus de Referência do Português Contemporâneo* (CLUL, Centro de Linguística da Universidade de Lisboa, 2008–2016) is provided with the *CQP* query language search engine, the reference corpora of English are focused on particular countries. The referential corpus of Spanish *CREA* (Real Academia Española, n.d.) does not allow for the usage of *CQL* or *CQP* and, logically, search by means of regular expression, and there is not any referential corpus for Ukrainian. However, given the fact that the team of the corpus *GRAK* team is making efforts to meet the referential criteria of the corpus and considering another advantage that interjections in this corpus are correctly tagged, we judge the corpus *GRAK* as the best source for achieving the goal set.

On the other hand, given the fact that onomatopoeias, especially occasionally created words, are likely to appear not only in fiction, but also in internet communication, we finally decided to use available *CQL* or *CQL* based referential of internet corpora of English, Portuguese, Spanish, and Ukrainian, particularly:

- English Internet Corpus from Leeds Collection of English Corpora (University of Leeds, 2022a), 190 million tokens.
- Spanish Internet Corpus from Leeds Collection of Internet Corpora (University of Leeds, 2022b), 145 million tokens.
- *CRPC*, Corpus de Referência do Português Contemporâneo (2008–2016), 411 million tokens.
- *GRAK*, General Regionally Annotated Corpus of Ukrainian (2017–2022), 1,476 million tokens.

3.4. Building queries

Corpus Query Language (CQL Guide), *Corpus Query Processor* (CQP, 2022) or alternative similar querying methods allow searching for given patterns based upon sequences of letters employing regular expression and corpus annotations.

Since many occasional onomatopoeias are not lemmatized nor annotated with particular tags in the corpus, their examples are to be chosen by the attribute *word*, which is aimed at selecting tokens with specific character sequences independently from their lexical and grammar properties, as shown in the example of Query 1:

(Query 1) **[word = "([bcdfghklmnprstvwxz] +) [aeiou] {1,2}**
([bcdfghklmnprstvwxz] +)-\1 [aeiou] {1,2}\2"]

The snippets inside the quotation marks composing the major part of Query 1 in *CQL* and *CQP* are processed as regular expressions. The regular expression between quotation marks is designed to match words that follow a consonant-vowel-consonant structure, where the first and last consonant combinations in syllables are the same, while the vowels may be either identical or different, e.g., *tic-tac, brum-brum*, etc. Let us now provide a detailed explanation regarding each part of the regular expression used in Query 1 in **Table 1**:

Finally, the entire expression matches the pattern of at least two consonant-vowel-consonant hyphenated similar syllables with the same consonants and varying vowels. More detailed explanations of the *CQL* syntax usage are accessible in the *Corpus Query Language Guide* (Sketch Engine, n.d.).

Query 1 is well suited for the corpora of English. However, in the case of applying a similar query to a language with different character sets (diacritics, Cyrillic, Greek, etc.), the characters inside the regular expressions are to be adapted to its alphabet, as we do for Portuguese, Spanish,

([bcdfghklmnprstvwxz]+)	matches the sequence of one or more consonants (represented by the character class inside the parentheses) and saves them as the first capturing group; here it stands for graphemes representing the consonant sounds of the language of corpus
[aeiou]{1,2}	matches one or two vowels; here it stands for other graphemes representing the vowel sounds of the language of corpus
([bcdfghklmnprstvwxz]+):	matches another sequence of one or more consonants and saves them as the second capturing group; here it stands for graphemes representing the consonant sounds of the language of corpus
–	matches a hyphen
\1:	backreferences the first capturing group, (i.e., it matches the consonants of the first capturing group, ensuring that they are repeated at the current position)
[aeiou]{1,2}:	matches one or two vowels; here it stands for graphemes representing the vowel sounds of the language of corpus
\2:	backreferences the second capturing group, ensuring that the same consonants captured in the second group are repeated here

Table 1: Description of functionality of parts of Query 1.

and Ukrainian. To apply the same query for the *Corpus de Referência do Português Contemporâneo* (CLUL, Centro de Linguística da Universidade de Lisboa., 2008–2016), Portuguese Language Corpus from the Leeds Collection of Internet Corpora (2022), we need to extend the character class with the diacritics (Query 2). Query 3 is respectively adapted for Spanish, and Query 4 for Ukrainian:

(Query 2) **[word="([bcdfghklmnprstvwxzç]+)[aieouáíéóúäâãéêõöô]{1,2} ([bcdfghklmnprstvwxzç]+)-\1[aieouáíéóúäâãéêõöô]{1,2}\2"]**

(Query 3) **[word="([bcdfghklmnprstvwxzçñ]+)[aieouáíéóúäâãéêõöô]{1,2} ([bcdfghklmnprstvwxzçñ]+)-\1[aieouáíéóúäâãéêõöô]{1,2}\2"]**

(Query 4) **[word="([бвгґджзклмнпрстфхцчшщ]+)[аіеоуяєю]{1,2} ([бвгґджзклмнпрстфхцчшщ]+)-\1[аіеоуяєю]{1,2}\2»]**

3.5. Validation of examples

We consider valid those examples that are interjectional onomatopoeias, i.e., sound-imitating words belonging to the class of interjections. Since the corpus query cannot delimit the sound-imitating words from those phonically motivated lexemes that are no longer interjections, but that could be qualified as such in the moment of creation, we considered those cases of

"etymological" onomatopoeias as valid examples (e.g.: *zigzag*, *criss-cross*, *flip-flops*, etc.). Although this decision may seem arbitrary, we aim to evaluate the queries' potentiality to match the necessary graphic patterns, rather than exploring their usability for distinguishing the evolution of the word meaning. At the same time, we discard from this survey other types of onomatopoeia expressed with nouns and verbs. For instance, the Query 7 (**Table 2**), among results, yields the words *murmur* and *barber*, that might be valid for other objectives. Whereas conventional onomatopoeias could be checked in the dictionaries, to judge the onomatopoeic function of occasional words, we use contextual analysis at the level of concordance line or paragraph.

4. Results and discussion

The queries created for extracting data can be qualified as models, since they represent a generalized schema suitable for the search of the extensive set of phenomena in question. The efficiency of a model can be measured differently by applying such parameters as *accuracy*, *precision*, *sensitivity* and other commonly used metrics.

In the case of using a model for previously unknown data without any established benchmark, it is impossible to calculate the sensitivity (also called *recall*), which is the number of retrieved true cases out of all the true cases in the population. Neither we can calculate the accuracy, which represents the number of true positive and true negative cases in relation to the entire number of cases in the dataset, as we cannot know the number of true negatives. In contrast, the precision demonstrates how many true cases appear in the selection, which is the parameter we aim to apply to perform a selection of as many as possible onomatopoeias out of a corpus, and it can be calculated on the data retrieved. For this reason, to roughly evaluate the validity of a query in terms of precision, we calculate the number of valid examples out of the first 100 examples in the concordance lines.

4.1. Retrieved data description

The representation of repeated syllables through grapheme level posits a series of questions such as syllables division, diphthongization and hiatus, unpronounced graphemes, and open and closed syllables. Nevertheless, some of these dilemmas can be overcome by assuming that onomatopoeic words are not created solely according to strict phonic patterns: sometimes the repetition may include one or several syllables (*meow*, *meow-meow*) and some onomatopoeias may present variants of graphic representations (*achoo*, *atchoo*, *achew*); therefore, there is probably little sense of rigorous compliance to the syllable divisions. Additional observations can shed light upon the fact that the vast majority of the onomatopoeias start with an initial consonant grapheme, end with another consonant grapheme, and, in separate cases, with a vowel. In other words, the main relevant feature to take into account for the corpus queries is to consider repetitions of sequence starting with consonant graphemes, followed by vowels and optionally ending with another consonant grapheme or a group of such graphemes.

Given that the hyphen can also be optional, the possible queries are to rely upon a four-member paradigm:

- repeated non-hyphenated closed syllables;
- repeated hyphenated open syllables;
- repeated hyphenated closed syllables;
- repeated non-hyphenated open syllables.

Hereafter, in **Tables 2, 3, 4** and **5** we expose the results yielded by the respective queries with the valid examples out of 100 first generated lines in the concordances; we indicate the number of repeated forms in parenthesis.

Type of syllables	**Query and extracted examples**	**Useful examples over 100**	**Overall results in the corpus**
Repeated hyphenated closed syllables	(Query 5) [word“([bcdfghklmn-prstvwxz]+)[aeiou]{1,2} ([bcdfghklmn-prstvwxz]+)-\1[aeiou]{1,2}\2”] **Valid examples:** *beep-beep, bling-bling, boing-boing, bon-bon, brrring-brrring, bump-bump, chit-chat, chop-chop, chow-chow, chug-chug, chun-chuan, clip-clop, cous-cous, criss-cross, der-der, dig-dug, ding-dong.*	61	956
Repeated hyphenated open syllables	(Query 6) [word=”([bcdfghklmnprstvwxz]+)[aeiou]{1,2}h?-\1[aeiou]{1,2}h?”] **Valid examples:** *bee-bee, beh-beh, bi-bi, bla-bla, blah-blah, boo-boo, cha-cha, chi-chi, choo-choo, coo-coo, da-da, do-dah, do-do, doo-dah, doo-doo, duh-duh, foo-foo, froo-froo, frou-frou, ga-ga, go-go.*	78	475
Repeated non-hyphen-ated closed syllables	(Query 7) [word=”([bcdfghklm-nprstvwxz]+)[aeiou]{1,2} ([bcdfghklmn-prstvwxz]+)\1[aeiou]{1,2}\2”] **Valid examples:** *boingboing, chinchin, hahhah, xiangxing.*	4	993
Repeated non-hyphen-ated open syllables	(Query 8) [word=”([bcdfghklmnprstvwxz]+)[aeiou] {1,2}h?\1[aeiou]{1,2}h?”] **Valid examples:** 0.	0	1000

Table 2: Results for the Leeds Collection of English Corpora (Internet Corpus).

Type of syllables	Query and extracted examples	Useful examples over 100	Overall results in the corpus
Repeated hyphenated closed syllables	(Query 9) [word="([bcdfghklmnprstvwxzç]+)[aieouáíéóúäâãéêõöô]{1,2}([bcdfghklmnprstvwxzç]+)-\1[aieouáíéóúäâãéêõöô]{1,2}\2"] **Valid examples:** *flic-flac(14), ping-pong (19), can-can (2), bip-bip, tan-tan, tim-tim(4), flics-flacs, zig-zag, zuns-zuns, den-den, hip-hop (28), tic-tac, tam-tam, tchim*-tchim (3).	78	1007
Repeated hyphenated open syllables	(Query 10) [word="([bcdfghklmnprstvwxzç]+)[aieouáí éóúäâãéêõöô]{1,2}h?-\1[aieouáíéóúäâãéêõöô]{1,2}h?"] **Valid examples:** *cri-cri, tsé-tsé (40), bla-bla (5), fru-fru, frou-frou, tau-tau, wha-wha (3), glu-glu (2), tai-tai, chi-chi, xi-xi.*	60	296
Repeated non-hyphenated closed syllables	(Query 11) [word="([bcdfghklmnprstvwxzç]+)[aieouáíéóú äâãéêõöô]{1,2}([bcdfghklmnprstvwxzç]+)\1[aieouáíéóúäâãéêõöô]{1,2}\2"] **Valid examples:** 0.	0	73,223
Repeated non-hyphenated open syllables	(Query 12) [word="([bcdfghklmnprstvwxz]+)[aieouáíéóúäâãéê õöô]{1,2}h?\1[aieouáíéóúäâãéêõöô]{1,2}h?"] **Valid examples:** 0.	0	322,820

Table 3: Results for the *CRPC*.

According to the results obtained, an outstanding fact that drew our attention was that for the queries 5, 6, 9, 10, 13, 14, 17 and 18, the rate of valid examples was much higher than those yielded by others. Many cases of onomatopoeias do not figure in the reference explanatory dictionaries. For example, out of 25 different onomatopoeias retrieved from the *CRPC* by queries 9 and 10, 17 do not appear as entries in the dictionary *Priberam* (2023) neither in bisyllabic or monosyllabic forms as sound-imitating lexemes: *cri-cri, frou-frou, wha-wha, glu-glu, tai-tai, chi-chi, flic-flac, can-can, tan-tan, tim-tim, flics-flacs, zig-zag, zuns-zuns, den-den, hip-hop, tic-tac, tam-tam*. This means that significant parts of the examples found are occasional sound-imitating words missing in lexicographic entries of modern dictionaries.

Type of syllables	Query and extracted examples	Useful examples over 100	Overall results in the corpus
Repeated hyphenated closed syllables	(Query 13) [word = "([bcdfghklmnprstvwxzç] +)[aieouáíéóúäâã éêõöüô]{1,2}([bcdfghklmnprstvwxzç] +)-\1[aieouáíéóúäâãéêõöüô]{1,2}\2"] **Valid examples:** *zig-zag, tut-tut, tun-tun, tic-tac, tap-tap, tan-tán, tam-tam, run-run, ruc-ruc, ris-ras, pon-pon, pis-pas, pin-pon, ping-pong, pim-pom, pil-pil, pill-pill, mish-mash, kin-kan, hip-hop, cric-cric, cric-crac, cous-cous, click-clack, chow-chow, chon-chon, chis-chas, chin-chin, chal-chal, can-can, bum-bum, boom-boom, bip-bip*.	91	304
Repeated hyphenated open syllables	(Query 14) [word = "([bcdfghklmnprstvwxzç] +)[aieouáíéóúä âãéêõöüô]{1,2}h?-\1[aieouáíéóúäâãéêõöüô]{1,2}h?"] **Valid examples:** *cu-cu, trau-trau, poo-poo, boo-boo, bu-bu, no-ni, reé-río, pro-prio, da-da, fío-fío, xie-xie, tsi-tsi, go-gó, feo-feo, blah-blah, bla-bla, frú-frú, pai-pai, re-re, deu-da, du-duá, cri-cri, cu-cú, pi-pi, wah-wah, cua-cua, tue-tue, bee-bee, tse-tse, fru-fru, ga-ga, ka-ke, tsé-tsé, pa-pa, pío-pío, fru-frú, mi-mi.*	45	62
Repeated non-hyphenated closed syllaba	(Query 15) [word = "([bcdfghklmnprstvwxzç] +)[aieouáíéóúäâãéêõ öüô]{1,2}([bcdfghklmnprstvwxzç] +)\1[aieouáíéóúäâãéêõöüô]{1,2}\2"] **Valid examples:** 0.	0	992
Repeated non-hyphenated open syllables	(Query 16) [word = "([bcdfghklmnprstvwxz] +)[aieouáíéóúäâãéêõöüô]{1,2}h?\1[aieouáíéóúäâãéêõöüô]{1,2}h?"] **Valid examples:** 0.	0	999

Table 4: Results for Spanish / Leeds Collection of Internet Corpora.

The impressive number of retrieved examples of repeated non-hyphenated open syllables observed in Portuguese and Ukrainian is likely due to a universal linguistic tendency. In the Leeds Corpora, the interface limited the maximum number of examples to 1,000, which is why we could not access the complete results.

Type of syllables	Query and extracted examples	Useful examples over 100	Overall results in the corpus
Repeated hyphenated closed syllables	(Query 17) [word = "([бвгґджзйклмнпрстфхцчшщ] +)[аіеоуяєю]{1,2}([бвгґджзклмнпрстфхцчшщ] +)-\1[аіеоуяєю]{1,2}\2?[аіеоуяєю]?.*"] **Valid examples:** *бен-бен-бен, бом-бом, брум-брумкаючи, ген-ген (6), гоп-гоп-ля, гур-гур, гуп-гуп, гав-гав, дзяв-дзяв (2), каґ-каґ, мур-мур (3), нюх-нюх, раз-раз, рох-рох, рох-рох-рох, тік-так, туж-тужб, тук-так-тук, тук-так-тук-так, туж-туж, туп-туп, човг-черх (3).*	32	35,202
Repeated hyphenated open syllables	(Query 18) [word = "([бвгґджзклмнпрстфхцчшщ] +)[аіеоуяєю]{1,2}-([бвгґджзклмнпрстфхцчшщ] +)[аіеоуяєю]{1,2}.*"] **Valid examples:** *го-го-го, гу-гу (2), гу-гу-гу (2), ку-ку, ха-ха, та-та, тра-та-та-та, ту-тум (4), тьфу-тьфу-тьфу (2), хі-хікання, ху-ху-ху, ха-ха (2), ха-ха-ха (2), шу-шу-шу (2).*	21	50,371
Repeated non-hyphenated closed syllables	(Query 19) [word = "([бвгґджзклмнпрстфхцчшщ] +)[аіеоуяєю]{1,2}([бвгґджзклмнпрстфхцчшщ] +)\1[аіеоуяєю]{1,2}\2?[аіеоуяєю]?.*"] **Valid examples:0**.	0	5,504,583
Repeated non-hyphenated open syllables	(Query 20) [word = "([бвгґджзклмнпрстфхцчшщ] +)[аіеоуяєю]{1,2}-([бвгґджзклмнпрстфхцчшщ] +)[аіеоуяєю]{1,2}.*"] **Valid examples:** *ду-ду, ха-ха-ха.*	2	432,002

Table 5: Results for the Corpus *GRAK*-16.

From **Tables 2–5** we can also observe that the more overall concordance lines are generated as per the query, the less specific the query focus of the onomatopoeias is.

4.2. Exploring markers' statistical significance

Table 6 integrates the number of specific results produced by the queries 5–20 in the four corpora excluding the types of queries that yielded 0 results.

Type of syllable	Valid examples over 100
Repeated hyphenated closed syllables (English)	61
Repeated hyphenated open syllables (English)	78
Repeated non-hyphenated closed syllables (English)	4
Repeated hyphenated closed syllables (Portuguese)	83
Repeated hyphenated open syllables (Portuguese)	60
Repeated hyphenated closed syllables (Spanish)	91
Repeated hyphenated open syllables (Spanish)	45
Repeated hyphenated closed syllables (Ukrainian)	32
Repeated hyphenated open syllables (Ukrainian)	22
Repeated non-hyphenated open syllables (Ukrainian)	2

Table 6: Integrated results: precision of patterns implemented in each query.

It is obvious that the repeated hyphenated closed syllables patterns yielded the most significant results among the four languages. To confirm that this is a tendency rather than an occasional combination of factors, we transposed the data as in **Table 7** and subjected it to an ANOVA test (single-factor), as we expected that in this experiment the only significant factor was the type of syllables.

	Repeated hyphenated closed syllables	Repeated hyphenated open syllables	Repeated non-hyphenated closed syllables	Repeated non-hyphenated open syllables
English	61	78	4	0
Portuguese	83	60	0	0
Spanish	91	45	0	0
Ukrainian	32	22	0	2
Average	68.67	51.25	1	0.5

Table 7: Number of useful results (over 100) as per each type of query.

Table 8 illustrates the results of the ANOVA test performed in Microsoft Excel.

Anova: Single Factor				
Groups	**Count**	**Sum**	**Average**	**Variance**
Column 1	4	267	66.75	697.5833
Column 2	4	205	51.25	562.25
Column 3	4	4	1	4
Column 4	4	2	0.5	1

ANOVA						
Source of variation	**SS**	**df**	**MS**	**F**	**P-value**	**F crit**
Between groups	14053.25	3	4684.417	14.81434	0.000245	3.490295
Within groups	3794.5	12	316.2083			
Total	17847.75	15				

Table 8: Anova test as per the data analysis in Microsoft Excel.

It is seen that the *p-value* (i.e., the probability that the achieved results are due to random coincidence) is equal to 0.000251. This value is far lower than the conventional 0.05 (i.e., the 5% threshold), which confirms that the data obtained are not due to chance, and the closed-syllable base onomatopoeias with repeated sounds turn out to be the most productive query pattern in the four observed languages, making the hyphen a robust onomatopoeic marker.

It is obvious that most hyphenated onomatopoeias do exist both in bisyllabic and monosyllabic forms, e.g., *bang-bang / bang, cling-cling / cling, ching-ching / ching, beep-beep / beep, plink-plink / plink*, and this property could be used at a further stage to optimize the extraction programmatically according to the following algorithm: if a pattern with repeated syllables recurrently occurs in a corpus, perform monosyllabic search for the syllabus used in the pattern.

4.3. Searching for interjectional onomatopoeias through POS-filter

Since interjections in the corpus *GRAK* have been correctly tagged, there is the possibility to perform the queries now with an additional pos-filter to limit the sampling exclusively to the cases of interjections, which increased the precision approximately twice. The queries 17 and 18 (**Table 5**) are now extended with this pos-filter. The results are shown in **Tables 9** and **10**, respectively.

Among the corpora involved in this survey, the only corpus with correctly tagged interjections is the corpus *GRAK*, hence we put to test only the selection in Ukrainian. The yielded results, once added the condition ([tag = ".*intj.*"]), are shown in **Table 9** and **Table 10**.

Query and extracted examples	**Language/ corpus**	**Useful examples over 100**	**Overall results**
(Query 21) [word = "([бвгґджзклмнпрстфхцчшщ] +)[аіеоуяєю]{1,2}([бвгґджзклмнпрстфхцчшщ] +)-\1[аіеоуяєю]{1,2}\2?[аіеоуяєю]?.*"&tag = ".*intj.*"] **Valid examples:** *мур-мур, рох-рох, тік-так, гур-гур, гуп-гуп, туп-туп, так-так, бом-бом, рох-рох-рох, дзяв-дзяв, гав-гав, клац-клац, цок-цок-цок-цок, кап-кап, тук-тук, клац-клац-клац, цур-цура, марш-марш, бум-бум, тук-тук-тук, цок-цок, няв-няв, туп-туп-туп, штовх-штовх, ціп-ціп-ціп, чах-чах-чах-чах, гам-гам, тік-тік, бом-бом-бом, так-так-так, цмок-цмок, свят-свят, свят-свят-свят, гав-гав-гав*	***CQL, GRAK* 16**	**92**	**3,513**

Table 9: Repeated hyphenated closed syllables (Ukrainian) with pos-filter.

Query and extracted examples	**Language/ Corpus**	**Use- ful examples over 100**	**Overall results**
(Query 22) [word = "([бвгґджзклмнпрстфхцчшщ] +)[аіеоуяєю]{1,2}-([бвгґджзклмнпрстфхцчшщ] +)[аіеоуяєю] {1,2}.*"&tag = ".*intj.*"] **Valid examples:** *ха-ха-ха, ху-ху-ху, го-го-го, ха-ха, ку-ку, ну-ну-ну, хе-хе-хе, го-го, ха-ха-ха-ха, га-га-га, ха-хо-хо-хо, хо-хо, ну-ну, ні-ні-ні, хе-хе, ква-ква, ме-ме-ме, цу-цу, ме-ме, та-та-та, ху-ху, ба-бах, ша-ша, хо-хо-хо-хо, тю-тю, го-го-го-го, ку-ку-рі-ку, хі-хі-хі, ту-ту, тю-тю-тю, ку-ку-рі-ку-у*	***CQL, GRAK* 16**	**91**	**5,284**

Table 10: Repeated hyphenated open syllables (Ukrainian) with pos-filter.

In fact, in many corpora the interjections appear tagged as nouns or adjectives. A serious challenge of modern corpus linguistics is the interjections recognition in transcribed corpora (Tellier et al., 2010):

> Among the seven consistent tagging errors presented above, some posed theoretical challenges due to their essentially pragmatic function and difficulty of fitting into a 'traditional' word class defined following morphosyntactic criteria. This is the case of pragmatic markers such as *well*, interjections such as *oh, ah*, and response forms such as *yes, no, okay, yeah, sure*. Other tagging errors only need a specific rule to help CLAWS4 disambiguate and assign the correct tag (Galiano, L. & Semeraro, A. Part-of-Speech and Pragmatic, 2023, p. 26).

The results of Query 22 (**Table 10**) confirm the demand for text corpora containing accurately tagged interjections. What catches our attention in **Table 9** and **Table 10** is the significantly superior performance of the query in the corpus *GRAK*, which is seemingly due to the presence of correct interjection annotation, whereas in the corpora of English, Spanish, and Portuguese utilized, the interjectional onomatopoeias are assigned tags of nouns or adjectives. This pos discrepancy makes it impossible to use the pos-attribute as a reliable feature for extraction.

4.4. Extraction from plain texts using regular expressions filter

The logical question arising from the observations of **Table 9** and **Table 10** is whether the query takes into account any linguistic feature. The corpora are textual databases provided with linguistic relevant data, but our queries performed to all the corpora except the corpus *GRAK* were based exclusively upon formal features such as grapheme combinations, with no relation to other linguistic properties. In fact, most queries include the attribute *word*. From the standpoint of corpus linguistics, *word* is a specific sequence of symbols separated by delimiters such as space. This means that the queries are intended to search particular sequences of letters by the regular expression syntax, which can also be successfully searched in many text editors with regular-expression engines, such as *Notepad++*, *Sublime edit*, or similar programs by applying the following query:

(Query 23) \b.*?([bcdfghjklmnprstvxzçñ]+)[aieouáíéóúäâãéêõöô]{1,2}([\w]+)-\1[aieou]{1,2}\2.*?\b

It can be observed that the regular expression from Query 23 was the same as in Query 22 (**Table 10**). As expected, the search within the novel *La sombra del águila* ('The Eagle's Shadow') (Pérez Reverte, 1993) yielded 100% of the valid results: *cling-clang* (seven times), *bang-bang* (three times), *zas-zas* (two times), and *ras-ras* (once). This result illustrates an ideal accuracy, but not in term of precision, since some valid cases of onomatopoeia might have been left out by the query. This text is rich in onomatopoeias since this novel narrates war events, which was the main reason to use it as additional empirical data.

The regular expression from the query used in **Table 11**, once applied to the text by Arturo Pérez Reverte, in contrast, did not yield purely onomatopoeic examples, however the found

matches were related to another sound imitating phenomena, since they all imitated stuttering speaking of a character: *vu-vuelto, lo-locos, lo-locos, su-suicidio, la-la, va-van, de-descuartizar, te-temo, po-posible, ma-malentendido, la-lapsus, he-hemos, po-polvo, ci-ciento, ma-mañana, su-suman, co-compañía, lui-la, pa-parece, se-setecientos, he-heridos, ci-cierto, du-duele, so-sombra, sa-sacrificio, ge-gesta, pe-perdido, nu-nuevecitas, de-demás, po-podéis, mi-mierda* (Pérez Reverte, 1993).

4.5. Three or more equal letters as a marker

Given the constraints of available space, we are unable to explore the marker based upon letter repletion in greater detail, but it is worth outlining some preliminary observations to be put to the test in further research. The consonant repetition at least three times as an onomatopoeic marker was earlier observed by Orrequia-Barea and Marín-Honor (2020, p. 53). Our observations suggest that additional markers' usage in queries to transcribed text corpora, such as exclamation marks, ellipses, or quotation marks, may significantly improve the results. We should also admit that both vowels and consonants are relevant in this pattern. The results are shown in **Table 11**.

Query and extracted examples	**Useful examples over 100**	**Overall results in the corpus**
• (Query 24) [word = "\w*([a-záíéóúäâãéêõöüô])\1\1\1\w*"] [word = "!|\.\.\.|\""] **Valid examples:** *Ohhhhh!, Ahhhhh!, Tssss, Ehhhh!, Zzzzzzzzzzzzzzzz, Hiiiiiiiii!, Hiiiiiiiiiii!, yeaaaargh!, Pfffff!, Uiiiiiiiimm, Uuuiiiiii!, zzzz, Prrriuuuuu, hurrrraaaaaaaaa!, vrrrr!, haaaaaa, goooooooo, ooooooo, oooooo, oooooo, ooooo, hurrrrraaaaaaaaa!, OOOhhhh!, Méééé!, dggggg, zzzz!, zzzzzz!, Uuuuu, Hmmmm!, vruuuumm!, Zzzzzzzzzp!, Zzzzzzzzp!, Aaaaah!, aaaaaaahs!, Psssst!, Aaaaaaah!, Aiiiii!, Zzzzzzzz, Bzzzz!, Aaaaaa, Aaaaatchim!, Oooootchim!, Trrrrrr, Trrrrri, trrrrru, booiiiii, Ummmm, Doooooiiiis, Iiiii, méeeeeeeh!, rrrrrt!, Ihhhhhh!*	**62**	**231**

Table 11: Repeated three or more letters followed by exclamation mark or ellipsis (*CRPC*, Portuguese).

The sound sources are easily deducible from the nearest context, as shown in the examples 7, 8 and 9:

(7) *Tenho de esperar que a máquina rebobine. **"Zzzzzzzzzzzzzzzz …"** – Idiota! (CRPC).*

(8) *O pior momento da campanha de Howard Dean veio depois da sua derrota no Iowa, quando proferiu um discurso em voz emocionada, que culminou num berro quase animalesco, **yeaaaargh!** (CRPC).*

(9) *Ainda assim a acção não perderá por completo a sua ligação ao universo dos quadradinhos, visto que as imagens reais se misturam com sequências de animação.* ***Vrrrrrrummmm****. Tac-tac-tac-tac-tac…* ***Uiiiiiiiiimm…*** *(CRPC)*

4.6. Other possible markers

Many onomatopoeic words are known to be neologisms, nonce or occasional words due to their creative nature. Therefore, they must appear in modern corpora as non-lemmatized wordforms, i.e., they must be stored in databases as words under unknown or empty lemmas. Hence, in the process of automatic lemmatization, the occasional onomatopoeias are not recognized as lexemes belonging to the given vocabulary and are assigned empty or "unknown" lemmas. *CQL* and *CQP* are provided with the possibility of searching for this kind of lemmas. In other words, rare lemmas, searched through the queries [lemma = ""], [lemma = "\|\|"] or [lemma = "unknown"] depending on conventions of a given corpus can increase the chance of finding exotic or occasional onomatopoeias.

Another promising formal feature to delve deeper into is the frequency factor, which can be also utilized as a marker: The transcribed onomatopoeias are likely to show lower frequency in the corpus in comparison to commonly used words, and *Sketch Engine* corpora provided with the *CQL* allow for applying the frequency as a separate filter in the query. The aforementioned and possibly other markers seem to be a promising perspective for research in further surveys.

5. Conclusion

Interjectional onomatopoeias are characterized by occasionality and wide variance; they are relatively rare in literature and are still out of the scope of the lexicographers of many languages or language combinations. While given word lists of conventional onomatopoeias provided in dictionaries are still quite limited, corpus queries allow for retrieving occasional sound-imitating lexemes on the basis of observed patterns, such as repetitions of graphemes and similar syllables sequences in combination with punctuation markers (hyphens, exclamation marks, ellipsis, quotes). The most fruitful proved to be the pattern of similar syllables. The best fitting patterns proved to be the letter combinations representing hyphenated similar (either open or closed) syllables, whose precision scored 66.67% for the closed syllables and 51.25% for the open syllables. These patterns proved efficient for the four involved languages, demonstrating similar tendencies. An ANOVA test proved that the revealed similarity was not due to chance. Thus, it is applicable to other languages.

It was observed that, in the case that the interjections in a corpus are correctly tagged as such, the precision increases approximately twice by including into the corpus query an additional pos-filter to rule out non-interjectional results. But, in lack of such, the regular expression syntax and the corpus query languages demonstrated similar efficiency for the closed hyphenated syllables

pattern. In contrast, for the open hyphenated syllables pattern, the regular expression in the searched text yielded 100% of a character's stuttering speech. Among the involved corpora, only in the Ukrainian corpus *GRAK* were the interjections consistently annotated with part of speech tags, and the precision of the query for interjectional onomatopoeias reached 92% for the hyphenated closed-syllables pattern and 91% for the hyphenated open-syllables pattern.

Although corpus queries, on the one hand, do not provide an exhaustive sample and, on the other hand, contain some redundant results, they nevertheless significantly speed up the search for illustrative examples that can be used for research and didactic purposes.

This study unveiled the perspectives that can be extended to monosyllabic variants of the extracted multisyllabic words. The conclusions obtained allow for further implementation of the pattern of three or more repeated letters along with punctuational markers, evaluating its precision, building, and exploring queries for extracting nominal, verbal onomatopoeias as well as other parts of speech tags with onomatopoeic characteristics and exploring the influence on the query precision of such additional factors as the token frequency or unknown lemma. Additionally, it is important to further develop methodological tools for elaborating principles of searching onomatopoeia in translation practice for denoting sound-imitating of particular phenomena, objects, and beings, and for working out criteria for establishing equivalent relations among onomatopoeias in different languages.

Competing Interests

The author has no competing interests to declare.

References

A bordo del Otto Neurath. (n.d.). *Pero, ¿Hay algo que sea la dialéctica?* [But, is there anything that might be dialectic?]. https://abordodelottoneurath.blogspot.com/2009/08/pero-hay-algo-que-sea-la-dialectica.html

Anderson Earl, R. (1998). *A grammar of iconism*. Madison, New Jersey: Fairleigh Dickinson University Press; London: Associated University Presses.

Assaneo, M., Nichols, J., & Trevisan, M. (2011). The Anatomy of Onomatopoeia. *PloS ONE*, *6*. DOI: https://doi.org/10.1371/journal.pone.0028317

Bidaud, S. (2022). Les onomatopées verbales du tcheque [Czech verbal onomatopeia]. *Studies about Languages*, *41*, 21–31. DOI: 10.5755/j01.sal.41.1.31330

Britannica. (2024). Onomatopoeia. https://www.britannica.com/topic/onomatopoeia

Casas-Tost, H. (2012). Translating onomatopoeia from Chinese into Spanish: A corpus-based analysis. *Perspectives Studies in Translatology*, *22*, 39–55. DOI: https://doi.org/10.1080/0907676X.2012.712144

Chamizo Babo, C. (n.d.). *Rapunzel*. https://www.eraumavezoutravez.com/rapunzel

CLUL, Centro de Linguística da Universidade de Lisboa. (2008–2016). *CRPC, Corpus de Referência do Português Contemporâneo* [Reference Corpus of Contemporary Portuguese]. http://gamma.clul.ul.pt/CQPweb/crpc/

Еґава, Х. & Кобелянська, О. (2016). *Японсько-український тематичний словник* ономатопеїчної лексики [Japanese-Ukrainian thematic dictionary of onomatopoeic vocabulary]. Kyiv: Dmytro Burago Publishing House.

Evert, S., & The CWB Development Team. (2022). *CQP Interface and Query Language Manual*. https://cwb.sourceforge.io/files/CQP_Tutorial/

Fundeu, Fundación del Español Urgente. (2011). *¡Tatatachán: 95 onomatopeyas!* [Tatatachán: 95 onomatopoeias]. https://www.fundeu.es/escribireninternet/tatatachan-95-onomatopeyas/

Galiano, L., & Semeraro, A. (2023). Part-of-Speech and Pragmatic Tagging of a Corpus of Film Dialogue: A Pilot Study. *Corpus Pragmatics*, *7*, 17–39. DOI: https://doi.org/10.1007/s41701-022-00132-9

GRAK. (2017–2022). *General Regionally Annotated Corpus of Ukrainian*. https://uacorpus.org/Kyiv/ua

Karamysheva, I.D. (2017). Contrastive Grammar of English and Ukrainian Languages. Vinnytsia: Nova Knyha Publishers.

Medvediv, A., & Dmytruk, A. (2019). Peculiarities of conveying the structural and semantic specificity of Japanese onomatopoeia in translation of texts of advertising character. *Research Trends in Modern Linguistics and Literature. Lutsk: Lesya Ukrainka Eastern European National University*, 2/2019, 77–94. DOI: https://doi.org/10.29038/2617-6696.2019.2.77.93

Meinard, M. E. M. (2022). *The Challenge of Defining Interjections and Onomatopoeias: A Contribution, Centered on Contemporary English*. [PhD Thesis, Université Lumiere Lyon 2]. https://www.academia.edu/67439120/The_Challenge_of_Defining_Interjections_and_Onomatopoeias_a_Contribution_Centered_on_Contemporary_English

Merriam-Webster Dictionary. (2024). Onomatopoeia. https://www.merriam-webster.com/dictionary/onomatopoeia?src=search-dict-box

Orrequia-Barea, A., & Marín-Honor, C. (2020). Building a parallel corpus of literary texts featuring onomatopoeias: ONPACOR. *Research in Corpus Linguistics*, *8*, 46–62. DOI: https://doi.org/10.32714/ricl.08.02.03

Pérez Reverte, A. (1993). *La sombra del águila* [The Shadow of the Eagle]. Madrid: Alfaguara.

Real Academia Española. (n. d.). Banco de datos (CREA). *Corpus de referencia del español actual* [Reference Corpus of Modern Spanish]. https://corpus.rae.es/creanet.html

Riera-Eures, M., & Sanjaume, M. M. (2010). *Diccionari d'onomatopeies i altres interjeccions: amb equivalències en anglès, espanyol i francés* [Dictionary of Onomatopoeias and other interjections with equivalents in English, Spanish and French]. Vic: Eumo.

Riondlearn. (2022). *Onomatopeia in Portuguese* [Onomatopoeia in Portuguese]. https://rioandlearn.com/onomatopoeia-in-portuguese/

Rodríguez Guzmán, J. (2011). Morfología de la onomatopeya. ¿Subclase de palabra subordinada a la interjección? [Morphology of Onomatopoeia: A Subclass of Word Subordinate to the Interjection?]. *Moenia*, *17*, 125–178. http://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=25534073

Round, E., & Kwon, N. (2015). Phonaesthemes in morphological theory. *Morphology*, *25*(1), 1–27.

RSVPLive. (2024). *Why hairdressers are the unsung heroes in our lives, writes Marguerite Kiely*. https://www.rsvplive.ie/life/hairdressers-unsung-heroes-lives-writes-14097868

Sketch Engine. (n.d.). *CQL Guide*: https://www.sketchengine.eu/documentation/cql-basics/

Sugahara, T. (2011). *Onomatopoeia in Spoken and Written English: Corpus- and Usage-based Analysis*. Hakkaido University. [Doctoral dissertation, Hokkaido University]. https://eprints.lib.hokudai.ac.jp/dspace/bitstream/2115/45138/1/Dissertation%20by%20Takashi%20SUGAHARA.pdf

Tellier, I., Eshkol, I., Taalab, S., & Prost, J.–P. (2010). POS-tagging for Oral Texts with CRF and Category Decomposition. *Research in Computing Science*, *46*, 79–90. https://hal.science/hal-00467951/documentddh

Universal POS Tags. *Universal Dependencies*. (2014–2024). https://universaldependencies.org/u/pos/

University of Leeds. (2022a). *Leeds Collection of English Corpora*. http://corpus.leeds.ac.uk/protected/query.html

University of Leeds. (2022b). *Leeds Collection of Internet Corpora*. http://corpus.leeds.ac.uk/internet.html

Vahidian Kamyar, T. (1996). *Farhange Namavaha dar Zbane Farsi* [Persian Onomatopoeia Dictionary]. Ferdowsi University of Mashhad Publication.

Yaqubi, M., Tahir, R., & Amini, M. (2018). Translation of Onomatopoeia: Somewhere between Equivalence and Function. *Studies in Linguistics and Literature*, *2*, 205–222. DOI: https://doi.org/10.22158/sll.v2n3p205

Yourdictionary. (2021). *Sound Words: Examples of Onomatopoeia*. https://www.yourdictionary.com/articles/sound-onomatopoeia-examples

Божко, І.С., & Кальніченко, А. (2023). Ономатопея як засіб експресивності в графічному романі: деякі нюанси перекладу [Onomatopoeias as an expressive means in graphic novel: some nuances of translation]. *Записки з романо-германської філології* [Notes on Romance and Germanic Philology], *2* (51), 30–41. DOI: https://doi.org/10.18524/2307-4604.2023.2(51).296818