knowledge

MDPI

Article

Investigating the Impacts of Misspellings in Patent Search by Combining Natural Language Tools and Rule-Based Approaches

Davide Russo [1,*], Christian Spreafico [1], Simone Avogadri [1] and Andrea Precorvi [2]

[1] Department of Management, Information and Production Engineering, University of Bergamo, 24044 Dalmine, Italy
[2] TRIX srl, 42015 Correggio, Italy
* Correspondence: davide.russo@unibg.it

Abstract: Among all sources of technical information, patent information is one of the richest and most comprehensive. Knowing how to search in this mass of documents is becoming increasingly crucial. However, many users have limited knowledge of patents and search strategies, so they must use intuitive, often approximate approaches that can lead to highly inaccurate searches and be time-consuming. To address this problem, there are tools that help expand queries to increase recall so as not to miss good documents, however, it remains an open problem dealing with misspellings-based strategies. Typically, the problem of the presence of misspellings in patent text is underestimated even by experts in the field, and there is no specific functionality to handle it in the tools available, both free and paid. The goal of the article is to raise awareness about the difficulties in making a proper patent strategy that also takes into account the possible presence of misspellings. It is important to know where we expect to find them and how much these may affect the final result. In particular, it is chosen to divide misspellings into categories, distinguishing between misspellings associated with a generic keyword or multiword from misspellings in acronyms, chemical formulas, names of applicants, inventors, or names of specific formulas or theorems. At least one example case is given for each category, showing when and how it may affect the result. Finally, an integrated approach combining word and contextual embedding models based on deep learning with a rule-based algorithm based on wild cards and truncation operators is suggested for correcting the query, automatically suggesting the most consistent misspellings, thus achieving a more accurate and reliable result.

Keywords: misspellings; typos; patent search; word embedding; contextual embedding; deep learning; natural language process

Citation: Russo, D.; Spreafico, C.; Avogadri, S.; Precorvi, A. Investigating the Impacts of Misspellings in Patent Search by Combining Natural Language Tools and Rule-Based Approaches. *Knowledge* **2022**, 2, 487–507. https://doi.org/10.3390/knowledge2030029

Academic Editor: Gwanggil Jeon

Received: 29 July 2022
Accepted: 29 August 2022
Published: 7 September 2022

1. Introduction

Patent information is one of the most interesting sources of technical knowledge. With its more than 130 million documents, it makes it possible to access information about inventions and technical developments from 1782 to today. Knowing how to search if there are relevant patents in this mass of documents is becoming increasingly crucial not only for those who want to file their patents but also for those who want to use this knowledge for other purposes (technological forecasting, competitors' intelligence, technology transfers, etc.).

The spread of patent searches is a relatively recent issue. Although patent databases have been collecting documents for a long time, patent searching coincides with the rapid expansion of the Internet. The development of the Web has made available, to everyone, an impressive number of documents that are accessible for free with search engines whose capabilities are still very small compared to the expectations of customers.

Despite the many advances made by international patent offices, consulting the various patent DB is a task for specialists skilled in the art. The less experienced today must live with all the difficulties that a patent text offers.

The patent searcher has created a query considering that every inventor has his/her style. For describing the same concept, it can adopt different expressions changing the level of detail (trellis frame vs steel tubular trellis frame); patentees prefer writing patents using very general terminology for extending claims validity (i.e., transport mean instead bicycle, bakery products for cupcakes). Sometimes vague, inconsistent, obsolete, or rarely used terminology is proposed for hiding patent content and making it difficult to search (hydrophilic surface instead of a contact lens). Even inaccurate terminology or neologisms are allowed to describe inventions in patents.

Handling the complexity of these aspects has been addressed in the literature, where we found several strategies and tools to help users. However, in patent strategies, there is a critical aspect that is largely underestimated. It deals with misspellings contained in the text. Although spell-checkers are becoming more widespread and increasingly effective, the presence of errors in patent texts is a fact of life. According to Intellevate Inc. [1], 98% of a sample of patents taken from the USPTO database contain errors, most of which are spelling errors.

The situation is not going to improve as with the advent of electronic filing of patent applications, the number of patent re-examination steps has been reduced. This has meant that the possibility of undetected spelling errors has greatly increased [2].

Once a patent is published, spelling errors contained in it are only removed by the USPTO upon request (U.S. Patent & Trademark Office, 2010). Regardless of whether they are unintentional, their presence can put a serious strain on search engines that are designed to look for the correct words and written text. The patent document is written in natural language, and therefore, each author may make spelling errors, translation errors, may misspell a term, automatic spelling checks are not perfect, not everyone who writes patents knows English perfectly, and most patents are not filed in English, but are translated (usually by automatic machine translation). Finally, there are problems with the conversion from one format to another. Each of these factors is responsible for transformations that change the text, often introducing errors or unwanted variants.

Most of patent providers do not offer dedicated services for misspellings' management and therefore, researching misspellings is left solely to the expertise of the practitioner. Today, the situation could change radically, since information retrieval techniques in patents are rapidly changing and new functionalities for error recognition can be implemented more easily. However, to develop these systems, it is necessary to understand what and how misspellings may be present in a patent and where.

The starting hypotheses of this study are the following:

- HP. 1: Misspellings can be present in all the parts of the text of a patent, e.g., description vs applicant or inventor name, etc.
- HP. 2: Different types of misspellings, contained in the same or different parts of a patent text, can influence the information retrieval from patents more or less severely.

To answer these hypotheses, this study proposes a systematic analysis and classification of the misspellings in patents by considering some patent searches having the following characteristics, differently from the previous contributions in the literature:

- Different search strategies (e.g., single or multi-words);
- The specific applicability for patents search;
- Different and random searched arguments (e.g., inventor, applicant, technology);
- The presence of different misspellings (in used queries), generated through tools for the misspellings generation;
- Validity for general and undomain application fields.

This paper is organized as follows. In Section 2, the state of the art about the previous ontologies for classifying misspellings in documentary searches and their limitations are

presented along with their limitations. In addition, a novel comprehensive classification of misspellings is introduced. In Section 3, an algorithm for searching patents with misspellings and investigating their impact is introduced. Section 4 presents the case studies in which the proposed algorithm has been applied. Then, the obtained results are presented in Section 5 and discussed in Section 6. The final section discusses conclusions and future research developments.

2. Literature Background about Misspellings Definition and Classification

A misspelling is defined as "a small mistake in a text made when it was typed or printed". Most misspellings involve simple duplication, omission, transposition, or substitution of a small number of characters. The distance between a single word and its misspellings is measured using the Levenshtein distance, that is the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one word into the other.

In the literature, several works propose ontologies of misspellings and study their impact on documentary research. The ontologies generally concern a certain field of application, such as the medical one (e.g. [3]) or the technological one (e.g. [4]). While, regarding the approach implemented to derive and classify the misspellings, different options have been considered. Several studies exploit dictionaries, such as wordnet (e.g. [5]), while others use the transliteration between one alphabet and another (e.g. [6]).

However, all these ontologies of misspellings are always dedicated to specific areas of application rather than to general purpose, and at the same time, they rarely deepen the specific problems of patent research, with the typical jargon of patents [1]. For this reason, a complete ontology of misspellings in support of patent research can only be constructed by combining the specific results of such ontologies in the literature, which has been done below, proposing an ontology that works for all types of documentary sources in the English language.

According to our classification, misspellings can be classified according to the origin of these errors (accidental or intentional), and further classifications can be identified for each of these categories as shown in Table 1.

Table 1. Misspellings' classification.

Accidental		**Voluntary**
Accidental Ignorance	**Accidental Typographic**	
Phonetically Plausible Misspelling Knok/knock	Thumbo, Twypo, Writo Bicylce, receive' as 'recieve	Typosquatting gogole.com (accessed on 28 July 2022)
Difficult Words (i.e., latin origins) Diarrhoea	Speako ate/eight	Neologisms Wake cup
Misuse and Orthographic errors than" and "then	Format conversion Universit?/università	Atomic misspellings prostate instead of prostrate
Compound (Hyphen or dash) email or e-mail	Transliteration of texts from non-latin alphabets Ko = co = cho	

2.1. Accidental Misspellings

In the category of accidental errors, ignorance errors were present. These derive, for example, from misspelling a word, and may depend on many causes as confusion due to the contracted form it's or its, spelling errors of words with similar phonetics (piece/peace, disc/disk), difficult words to spell such as diarrhea or pyrolysis, or from the misuse of the space everyday vs every day and more. The writer may lack the knowledge

or understanding of the phoneme–grapheme correspondence system in English, and/or the patterns of syllables written in English and their assembly into longer words and compounds. These kinds of errors are systematic; the user will keep getting it wrong every time he writes that term. On the contrary, occasional errors are unintentional mistakes that in past were mainly due to mechanical failure or slips of the hand or finger. They are much more frequent now because they are due to typing on touch-screen devices, hence the term Thumbo related to the expression all thumbs. Similarly, the Twypos are the errors in the tweets, (tweet + typo). Writo is a "typo" made by handwriting (an oversight due to inattention) and Speako, a similar mistake you make when speaking instead, especially when dictating to a speech recognition system. The most frequent misspellings are the omission of a letter (carbon/carbn), the addition of a letter (carbon/carbone), a single letter instead of a double one (address) and vice versa (carbonn), the substitution of a letter (calbon), the interchange of two adjacent letters (carbno), wrong punctuation or a space.

Accidental errors also include errors that do not result from mistyping text but are errors that result from the transformation of text during visualization or storage, zipping, cleaning. The best known are errors in the conversion of one text from "character-set" to another. On the DB the word is spelled correctly (i.e., Università) while on the web page the accented letters are replaced with "?" (i.e., University?). The terms most affected by this are accented characters and multiword strokes.

Accidental errors can also be produced by a machine converter transliterating a text from one alphabet into another one, for example Cyrillic into the Latin alphabet. (e.g., кириллица Cyrillic alphabet when translated into Latin ISO 9 is kirillica, in English is kirillitsa). For instance, ref. [7,8] use a strictly word-based approach and only handle the correction of out-of-vocabulary words into in-vocabulary words, while [9] use a noisy channel model conversion that consists of converting the input phonetic strings provided by the user into the appropriate word string using ideographic characters.

Another important aspect to be considered in making patent queries is to manage compounds. Basically, compound is a word that consists of two or more parts that work together as a unit to express a specific concept. They are written in one of these three ways: solid (e.g., teapot), hyphenated (player-manager), or open (washing machine). Choosing the right form to describe the compound represents one of the thorniest style problems that writers encounter. One need only think about compound nouns written as one or two words such as Iceland, shopkeeper, website, car park, or the following compound verbs (go-kart, breakwater, runway, check-in), and compound adjectives hyphenated—absent-minded and not hyphenated—small talk, greenhouse. Sometimes all three forms are accepted: lifestyle, life–style, or life style. Without a universally recognized precise grammatical rule, it is to be expected that there will be many patent writers who use incorrect forms. The possibility of error is even greater in the case of prefixed (such as anti-, non-, pre-, post-, re-, super-), suffixed (as -er, -ism, -ist, -less, -ful, -ness), and combining form compounds (mini-, macro-, pseudo-, -graphy, -logy).

The misuse of hyphens or dashes represents another source of involuntary errors. A hyphen is a punctuation mark used to divide or to compound two or more words or numbers together. A dash separates words into parenthetical statements. The two are sometimes confused because they look so similar, but their usage is different. Hyphens are not separated by spaces, while a dash has a space on either side. For example, e-bike, a pick-me-up, mother-in-law, good-hearted. A dash can be also substituted by a minus sign or an underscore.

2.2. Voluntary Misspellings

Certain misspellings are occasionally used deliberately to create neologisms or prevent someone from being easily found, or for humorous purposes. When a typo results in a correctly spelled word that is different from the intended one is called an "atomic typo". It is used to play word games and it is insidious because since it is spelled correctly, a simple spellchecker cannot find the mistake.

Typosquatting is a form of cybersquatting that relies on typographical errors made by users on the Internet. Typically, a cybersquatter will register a likely typo of the address of a frequently accessed website or deliberately introduce misspellings into a web page, or into its metadata, hoping to receive traffic when internet users enter these misspellings into online search engines. An example of this is gogole.com instead of google.com (accessed on 28 July 2022).

2.3. Misspellings Generation

Most methods and applications for handling misspellings were created to identify and correct misspellings in order to write text in its correct form. In synthesis, the methods used to correct context-sensitive spelling errors can be separated into three categories: rule-based, statistical, and deep learning-based methods [10].

2.3.1. Rule-Based

Under this category, numerous methods appear that try their hand at systematically removing, adding, or swapping one or more letters at a time from a source term. As Levenshtein's distance increases, the number of misspellings that can be generated grows exponentially. One way to limit the generation of Writos to those most likely to be generated is the keyboard adjacency rule, also used by Google's misspelling checker. When you hit the wrong key, Google would look to nearby keys to see which was the most likely one you were aiming for. This general concept was applied to all new misspellings, going through nearby letter replacements until a popular replacement term was found (useful in finding atomic misspellings). Other rules are: SoftTfIdf, a similarity metric to find correspondence in names [11], Soundex, which relies on similarity hashing [12], SmartSpell, a phonetic production approach that computes the likelihood of a misspelling [13], PatBase and FreePatentsOnline to handle near-duplicates in assignee name spelling correction dictionaries (e.g., ASpell).

2.3.2. Statistical Method

The statistical method is particularly suitable for spelling errors that have low repeatability. Usually, context analysis is the key to understand what the user is looking for regardless of the type of spelling error made, even if the error has never been seen before. A famous method for generating candidate words using contextual information is called 3-g [14].

Industry and universities have been interested in the practical development of solutions based on statistical methods, as evidenced by the many patents proposed on the subject. Consider, for example, these companies: Microsoft (US20050216253 A1), Xerox (US20130030787 A1), Covera Health (EP3956900 A1), Cimpress Schweiz (EP2705443 A1), Datacloud Technologies (US20020095448 A1), Ternarylogic (US20090045988 A1), Nully (KR20110005932); and the Universities: Pusan (KR20150007647), King Abdulaziz City for Science & Technology (EP2653982 A1), Kunming University of Science & Technology (CN106294315 A).

2.3.3. Deep Learning-Based Method

According to [10], spelling error correction techniques that employ language models are divided into four main categories: word embedding information-based, contextual embedding information-based, auto-regressive language model-based, and auto-encoding language model-based correction techniques.

In particular, word embedding models are a type of neural networks that maps words to a vector space of fixed dimension, they learn vector positions through training on the big textual corpus. Word2Vec, GloVe, FastAI are open-source word embedding models transforming words into vectors. They encode the meaning of words into short, dense vectors (word embeddings) that contextualize the meaning of words in any given corpus by looking at the words surrounding that word in the corpus. Because of this, they can be

used in so many contexts and applications like question answering, information retrieval, machine translation, language modelling, and misspelling checkers. According to W2Vec, words that exist in similar contexts in sentences are mapped to the same vector space. This means that words with similar neighboring/surrounding/context words in a corpus have similar vectors (with high cosine similarity). In this way, it is possible to find misspellings (wrong words related to the correct words as working in the same context as the correct words) by computing the distance among their vectors.

Among the tools used for this purpose, neural networks are the most diffused in many years (e.g. [15–17]). These tools are particularly valuable since they allow the detection of misspellings without pre-categorizing them through external supervision (e.g. [18,19]).

Companies have also become interested in deep learning and are patenting several methods based on it, including self-learning contextual spell correctors (e.g., Microsoft: US8176419B2), misspelling identification in domain names (e.g., US10380210B1), domains based on popularity (e.g., US20110258237A1), corrections of orthographic errors (e.g., CN110457688A, WO2021235968A1, US10115055B2).

3. Methodology

To answer the starting hypotheses, an algorithm was proposed to demonstrate the impact of misspellings in information retrieval from patents by comparing the results obtained from the application of a traditional algorithm for patent search and a variant including misspellings in different case studies.

The proposed algorithm is divided into the following three steps.

- Step 1 (selecting the patent database) regards the selection of the patent database within which the misspellings must be searched.
- Step 2 (classical patent search) defines the queries with which to search for the patents within a certain topic, containing grammatically correct keywords that define this topic. In this case, the goal is to count the number of patents that contain the correctly searched and written keywords.
- Step 3 (patent search including misspellings) defines the queries with which to search for patents that contain the same information but might be written incorrectly. For this reason, the keywords contained in these queries are the combinations of all the incorrect ways in which such information can be written. Within step 3, the misspellings are generated through a combination approach between natural language tools and rule-based approaches.

Therefore, the application of the proposed algorithm to some selected case studies is possible to evaluate the impact of misspellings on patent research. This is done by comparing the number of patents not containing the misspellings, retrieved through the application of step 2, and the patents containing the misspellings, retrieved through the application of step 3, for each case study.

Figure 1 graphically represents the proposed algorithm to evaluate the impact of the misspellings in a patent search.

In the following paragraphs, the steps of the proposed method are explained in detail.

3.1. Step 1—Selecting the Patent Database

The search was conducted on the entire patent database using the Fampat database. The Fampat Collection is a comprehensive family coverage of worldwide patent publications published by more than 100 patent authorities. In Fampat, a single-family record combines all publication stages of each family member. Search numbers always refer to the number of patent families. This database was preferred over others because it allows searching within only the parts of documents belonging to the same family written in English. Generally, the fields chosen were English title, English claims, and English description (indicated with eti/eclm/edesc).

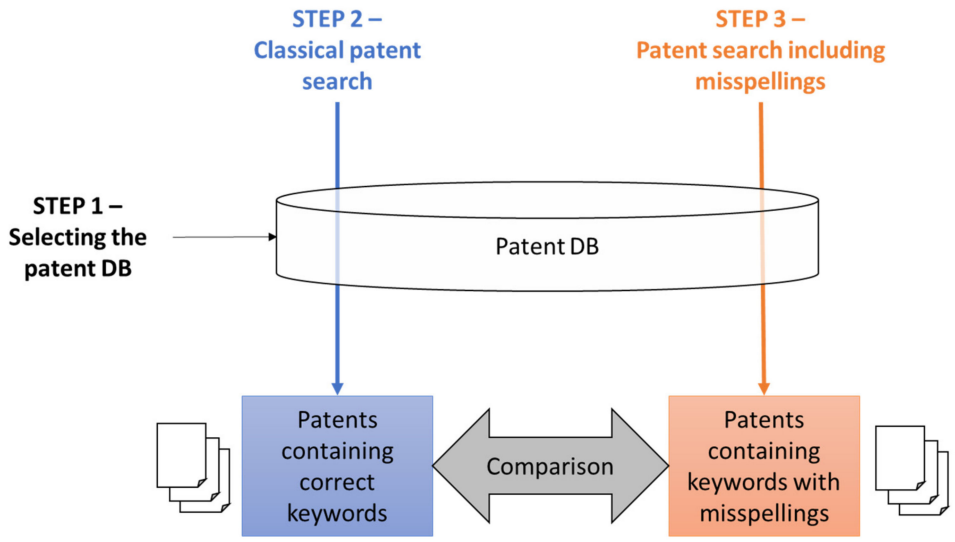


Figure 1. Graphical representation of the proposed algorithm to evaluate the impact of the misspellings in a patent search.

3.2. Step 2—Classical Patent Search

Since in the literature there is no search strategy widely recognized as traditional, for this particular scope we take into account two patent sources (cpc and semantic expansion) combined to build the pool. For the sake of simplicity, the traditional patent search algorithm presented in this article has been conceived to work with the English language, since the knowledge base of synonyms contains only English terms.

The basic schema for a single keyword was conducted mainly following this pattern: "((Keyword)/eti/eclm/edesc OR (English Synonyms of Keyword)/eti/eclm/edesc) AND (CPC#1 or CPC#n)/class".

Meanwhile, in case the target keyword was a compound, the reference query was: "(((keyword #1 OR Eng. Synonyms of K#1) distance (keyword #2 OR Synonyms of K#2)) //eti/eclm/edesc AND (CPC#1 or CPC#n)".

According to this method, we set one or more keywords representing the concept we are looking for, and all their synonyms from the English dictionary to improve the recall. In some cases, in order to improve precision, a CPC list is added to the query. Other strategies, including citation expansion, usually useful when initial searches provide only a few patents, were not necessary in our case.

3.3. Step 3—Patent Search including Misspellings

To show the influence of misspellings, the traditional search query was taken as a reference and modified by replacing the starting keyword with an equivalent misspelled keyword while leaving the rest of the query unchanged. In this way, it was possible to assess how many patents contain errors by weighing the results for each typo. The process was then repeated for all misspellings. Only the misspellings that gave results greater than zero were included in the table of the results. The generation of misspellings was limited in number only to those words that deviated from the source term by minor changes (Levenshtein's maximum distance equal to two).

The reference query to calculate misspellings of a single keyword was built substituting the keyword with his typo as follows: "((TYPO#i of the keyword) /eti/eclm/edesc OR (English Synonyms of Keyword)/eti/eclm/edesc) AND (CPC#1 or CPC#n)/class". While for compound, only one keyword at once was substituted by his typo, as follows: "(((TYPO#i of keyword OR Eng. Synonyms of K1) distance (keyword #2 OR Synonyms of K#2)) //eti/eclm/edesc AND (CPC#1 or CPC#n)".

To generate the misspellings (i.e., "TYPO" in the queries reported before), a combinate approach between natural language tools and rule-based approaches was used.

For the natural language approach, a variety of different word embedding tools has been tested GloVe Stanford University [20], fasttext [21], roBerta [22] and Bart (Facebook) [23], ELMo (Allen NLP) [24], GPT (OPEN AI) [25], Bert (Google AI) [26], XLNet (Google Brain) [27]. Being language agnostic, all word embedding tools suggest related words regardless of how they are spelled. Usually, they are used to find a thesaurus of terms for query expansion, in this case, word embedding can also identify misspelled related words regardless of how they are written or what characters they contain inside. Since these approaches work on mathematical computation, embedding words tends to work best with those misspelled words appearing repeatedly as substitutes. Unlike a parser that goes crazy with complex sentences [28], where the text is poorly written, the sentences are in ungrammatical English or if the constructions are too complicated, these models also work well with them. One parameter that is used to handle this is window size, which is how many words you consider before and after the i-th word. One of the key aspects in managing these patterns is setting the threshold of section of how many times a word should appear in the text. In case you want to consider even very low-frequency misspellings you have to deal with a model that becomes huge.

Hence the idea of combining two techniques, combining word and context embedding models with a series of rule-based algorithms. These will help to improve the performance, for example in cases of isolated misspellings, such as in the search by the applicant or by the inventor, in which there is no sentence to reconstruct the role of a word in a context. They are also useful in the case of atomic typos capable of distorting the result of the parsers.

The set of considered misspellings has been designed to demonstrate that the possibility of making a mistake is not the same for every word but there are scenarios with greater impact (think of long words composed of many syllables, with diphthongs capable of generating double, perhaps of Latin origin). In the alternative, think of the complexity of writing a chemical formula could be considered. These situations require more attention on the part of those preparing the search query.

For this purpose, a list of the most critical situations, built on the most frequent mistakes made during the writing phase, has been proposed. In many of these cases, it is possible to create rules from English grammar and use the truncation operators to cover the options that you think are easier to get wrong. For example, most of the problems on a multiword search, but not only that, can be overcome thanks to the introduction of truncation symbols (if the search engine provides them). Truncations are symbols that replace one or more characters, making it possible to search for different variants of a term. For truncation to work properly, the abbreviated term must contain at least three characters. Among these for example there are the following symbols:

- (+) right Truncation replaces any number of characters at the end of a term (bicycle+);
- (?) Truncation replaces zero or one character (bicycle?);
- (#) Truncation replaces exactly one character (b#c#cle);
- (_) The underscore allows for simultaneous searching of terms that may be written as one or two words. It will also retrieve results where there is a hyphen between terms, and it can also be used in chemical formulas.

Each engine has its syntax, and these symbols can be replaced by other special characters, but the basic functionality remains the same.

In Table 2 the list of suggested rule-based indications is reported.

Table 2. Suggested rule-based indications (where "+" = replaces any number of characters, "_" = allows for simultaneous searching of terms that may be written as one or two (ordered) words, "?" = replaces zero or one character, "#" = replaces exactly one character).

Example	Goal Terms	Rules	Example of Results
Compounds	Anglo Saxon	Anglo_saxon	Anglo-Saxon
		Anglo 0W saxon	Anglo Saxon
		Anglo?saxon	Anglossaxon
Common inflectional suffixes on base words (-s, -ed, -ing, -er, -est)	Bikes	Bike?	Bike, Bikes, Biker
	Measured Measuring Measurer	Measur+	Measure, Measures, Measuring, Measured, Measurement
		Measur??	Measure, Measures, Measured, Measuring (n.f.), Measurement (n.f.)
		Measure?	Measured, Measures, Measuring (n.f.), Measurement (n.f.), Measure (n.f.)
		Measuri??	Measuring, Measure (n.f.), Measures (n.f.), Measured (n.f.), Measurement (n.f.)
		Measur???	Measure, Measures, Measuring, Measured, Measurement (n.f.)
Most common prefixes: • closed and vowel-r syllables: non, ex, con, per, mal • open syllables: bi, co, di, o, pro, tri, twi, pre • two syllables: super, circum, intra, contra, counter, extra, intro, multi, ultra	Not critical	Not_critical	Not critical, Not-critical, Notcritical
	Bi-Phase	Bi_phase?	Biphase, Bi phase, Bi-phase
	Multi-Object	Multi_object	Multi object, Multiobject, Multi-object
		Multi+_object	Multi object, Multiobject, Multi-object, Multiple object, Multipart object, Multiplying object, Multipurpose object, Multipart object
More prefixes (fore, inter, trans, over, sub, semi, anti, mid, ex, post)	All prefix	Over+	Over, Overs, Over tube, Overwrapping, Overstrike, Overvoltage, Overlapping
		Over##+	Overwrapping, Overstrike, Overvoltage, Overlapping, Over, Overs, Over tube
	Prefix + known word	Fore_casting	Forecasting, Fore casting, Fore-casting
		Fore_cast+	Forecasting, Fore casting, Fore-casting, Forecast, Forecasted, Forecasts, Forecastable, Forecastiong, Forecastting
		Over_pressure	Over pressure, Over-pressure, Overpressure
		Over+_pressure	Over pressure, Over-pressure, Overpressure, Overwrapped pressure, Overall pressure, Overly pressure
Common suffixes beginning with a consonant (-ly, -ful, -ment, -hood, -less, -ness)	-ful	+ful	Ful, Useful, Powerful, Harmful
		+#ful	Ful, Useful, Useful, Powerful, Harmful
	-ful variation for known word	Useful	Useful
		Useful+	Useful, Usefull, Usefulness, Usefully, Usefulness, Usefullness
Suffixes with ti, ci, si (tion, sion, tious, sious, cial, tial)	-tious	+tious	Infectious, Sedimentitious, Nutritious, Surreptitious
	-tious variation for known word	Nutritio?s	Nutritious, Nutritios
		Nutriti?us	Nutritius, Nutritious
		Nutrit#ous	Nutrituous, Nutritious

Table 2. *Cont.*

Example	Goal Terms	Rules	Example of Results
In (immigrate, illegal, irregular) ad (address, approach, aggressive) ob (obstruct, opportunity) sub (subtract, suppose, surround) com (commit, collide, corrode) dis (dissuade, difference)	Double letter	Im?igrate	Immigrate, Imigrate
		Il?egal	Illegal, Ilegal
		Dif?erence	Difference, Diference
	Double letter replaced	Su#pose	Suppose, Subpose
Graphemes unique to Greek-based words ch = /k/ (chorus, monochrome) y = [I] or [#j] (dyslexia, cytoplasm) ph = /f/ (phonology, grapheme) x = /z/ (xylophone)	CH → K	Mono#?rome	Monochrome, Monokrome
	Y → I/J	Bic#cle	Bicycle, Bicicle
	PH → F	Gra#?ene	Graphene, Graphene,
	X → Z	#ylophone	Xylophone, Zylophone
Silent letter spellings rh (rheumatoid) mn (mnemonic) pt (pterodactyl)	Rh, ps, pn, mn, pt	R?eumatoid	Reumatoid, Rheumatoid
		M?emonic	Mnemonic, Memonic
Connectives that join the root and suffix i (menial, lenient, anxious) and u (superfluous, disingenuous, factual)	Presence or assence of connection	Superfl?ous	Superfluous, Superfluous

4. Case Studies

The criterion by which the case studies were chosen is inspired by the different types of research that a practitioner experiences. In the case of a prior art search or a freedom-to-operate search, it is important to set up a strategy that maximizes recall to be sure not to miss even one relevant result. Thus, in this case, it is important to know the absolute number of patents that contain a typo and could therefore escape a traditional search. To simulate this type of search, we started with a search for an object such as "bicycle" expressed in English as a single word and having a maximum of three syllables to limit the number of potential errors. The search was then replicated on a compound word such as "carbon dioxide". This target is a combination of two words and can be also expressed in its alphanumeric form "CO_2". Then the case of a compound containing a proper name such as "Brayton cycle" was chosen. The last exemplary case was "supercritical fluid" a complex compound, resulting in a combination of a noun and a compound adjective, where the adjective is itself formed by a compound of two adjectives.

To simulate a competitor analysis search, we thought of looking for misspellings related to the name of a famous company like "BOSCH". It is a short name with only five letters to limit the potential number of errors. The main risk of this search is to find atomic misspellings and wrong names corresponding to real companies different from Bosch. To be sure that the result always refers to "Robert Bosch gmbh" we decide to take into account only patents with misspelled Bosch as applicant together with well-written "Robert Bosch" as co-applicant.

There can also be errors in the designation of CPCs or IPCs but in this case, the typographical error is negligible compared to the much more pronounced error of the corresponding class assignment as explicated in [29].

The last search was designed for a dual purpose, to simulate the search by the inventor and simultaneously show the problems of transliteration. Among all the alphabets, the Cyrillic alphabet was chosen. Starting with the patents of Mr. "Sergei Ikovenko", we compared the search results with his name written in English and the name written in [30]. Cyrillic ("Яковенко"). The right name was taken from his institutional personal web page, where it was also possible to deduce from his curriculum vitae how many patents he had filed as an inventor.

Table 3 reports the queries with correct keywords and keywords with misspellings used in the considered case studies.

Table 3. Queries with correct keywords and keywords with misspellings used in the considered case studies.

Case Studies	Used Queries with Correct Keywords	Used Queries with Keywords with Misspellings
BRAYTON CYCLE	(BRAYTON+ 5D (CYCL+ OR CICLE+ OR THERM+_DYNAM+))/ETI/ECLM/DESC	((MISSPELLINGS OF BRAYTON+) 5D (CYCL+ OR CICLE+ OR THERM+_DYNAM+))/ETI/ECLM/DESC
BICYCLE OR BICYCLES	(BICYCLE OR BICYCLES)/TI/CLMS/DESC AND (B62+)/CPC/IPC	((MISSPELLINGS OF BICYCLE OR BICYCLES) 5D (CYCL+ OR CICLE+ OR THERM+_DYNAM+))/ETI/ECLM/DESC
BOSCH	BOSCH/PA/OPA	BOSCH/PA/OPA AND (ROBERT 1W MISSPELLINGS OF BOSCH)
CO_2	(CO_2)/ETI/ECLM/DESC	(MISSPELLINGS OF CO_2)/ETI/ECLM/DESC
CARBON DIOXIDE	(CARBON 0W DIOXIDE)/ETI/ECLM/DESC	(MISSPELLINGS OF (CARBON 0W DIOXIDE))/ETI/ECLM/DESC
SUPERCRITICAL (fluid)	(SUPERCRITICAL)/ETI/ECLM/DESC	(MISSPELLINGS OF SUPERCRITICAL)/ETI/ECLM/DESC
SERGEI ALEXANDROVICH IKOVENKO	(ЯКОВЕНКО AND (СЕРГЕЙ OR АЛЕКСАНДРОВИЧ))/IN/OIN/INH/INV	(MISSPELLINGS OF IKOVENKO) AND ((MISSPELLINGS OF SERGEI) OR (MISSPELLINGS OF ALEXANDROVICH)))/IN/OIN/INH/INV

The selection of this research does not presume to be exhaustive. Our goal was to show some quantitative results in order to be able to make considerations about the usefulness of introducing misspellings into the source query.

5. Results

5.1. BRAYTON CYCLE Search

Table 4 reports the results, in terms of patents, obtained using the queries with the correct keyword (i.e., "(brayton+ 5d (cycl+ or cicle+ or therm+_dynam+))/eti/eclm/desc") and with keywords containing by misspellings.

Table 4. Keywords (correct and with misspellings) used to search patents about BRAYTON CYCLE.

Keywords		Results (N° Patents)
Correct keyword	BRAYTON CYCLE	5395
Keywords with misspellings	BRAITON CYCLE	10
	BRAY-TON CYCLE	33
	BRYTON CYCLE	57
	BRIGHTON CYCLE	96
	BRETTON CYCLE	99
	BRITTON CYCLE	4
	BREE TON CYCLE	1

Although the numbers are not high, we are still talking about a percentage of about 5%. Furthermore, it is amazing to see how important companies such as General Electric, Shell, Hitachi, Politecnico di Milano can make this mistake (see Table 5).

Table 5. Misspellings contained in patents of notorious applicants.

Contained Misspellings	Patent Number and Applicant	Sentences with Misspellings
BRAITON CYCLES	WO2012/114367 HITACHI	The heat cycle of the gas turbine power generation system basically follows Braiton cycles, and the thermal efficiency is determined by the air compression ratio.
BRYTON	WO2022/117228 NUOVO PIGNONE TECNOLOGIE	The thermodynamic system 35 may include an open thermodynamic cycle, such as a Bryton cycle, using a gas turbine engine.
BRAITON CYCLE	ES2643558 SHELL	Any suitable liquefaction cycle known in the art may be used, including the Claude cycle, the Braiton cycle, the Joule Thompson cycle, and any modifications or combinations thereof.
BRYTON CYCLE	WO2019/194670 HYUNDAI HEAVY INDUSTRY	In the present embodiment is provided with a refrigerant heat (1275) exchanger N2 Bryton cycle the refrigerant supply portion (127) may be provided but, in any shape including a refrigerant heat (1275) exchanger is the first embodiment.
BRYTON CYCLE	WO2014/087344 ENEL INGEGNERIA & RICERCA-POLITECNICO DI MILANO	In this further secondary exchanger, the gas transfers the heat amount necessary to feed a Bryton cycle for micro-generation in an appropriate section of the plant by heating of air under pressure.
BRIGHTON	RU2719413 GENERAL ELECTRIC	Figure 1 depicts schematic diagram of the traditional system with Brighton's enclosed regenerative cycle for electricity generation;
BRIGHTON	RU2018129741 NUOVO PIGNONE TECNOLOGIE	Floating heat can be converted into useful energy through various thermal engines using thermodynamic cycles such as Renkin steam cycles, organic cycles of Renkin or Brighton, CO cycles [2] or other energy cycles.

5.2. BICYCLE Search

With a three-syllable word, the combinations of misspellings are innumerable. We have generated more than 100 and over 20 spelling variants have been manually identified that produce non-zero results. Since some words could be atomic misspellings such as bicile or bi cycle that are used in the biomedical or chemical fields, a filter on the B62 + class has been added.

Table 6 reports the results, in terms of patents, obtained using the queries with the correct keyword (i.e., "(BICYCLE)/ETI/ECLMS/EDESC AND (B62+)/CPC/IPC") and with the keywords with the misspellings.

Compared to the final result, patents containing bicycle typos are less than 1%. Moreover, it is not necessarily the case that a patent contains only wrong ways of bicycle; in fact, most of the time bicycle is spelled correctly. In this case, a keyword search using only correct keywords would still have retrieved the patent containing typos among the results. The fact remains that within that 1% patents appear where there are only typos, and for a prior art analysis this represents a very high risk of overlooking potentially important patents. Furthermore, it is incredible how companies have been found that make the wrong way to write bicycle in the applicant field and that they can write it wrong also in the title (see Table 7).

5.3. BOSCH Search

Table 8 reports the results, in terms of patents, obtained using the queries with the correct keyword (i.e., "(BOSCH/PA/OPA") and with the keywords with the typos, used in the query "(BOSCH AND (ROBERT 1W TYPO))/PA/OPA".

Table 6. Keywords (correct and with misspellings) used to search patents about BICYCLE.

Keywords		**Results (N° Patents)**
Correct keyword	BICYCLE	188,699
Keywords with misspellings providing more patents (>100 each)	BICY	823
	BICYLE	417
	BYCYCLE	132
	BI-CYCLE	128
	BI CYCLE	128
	BICI	126
	BYCICLE	117
	ABICYCLE	109
Other keywords with misspellings	BICICLE	78
	BI CICLE	0
	BI-CICLE	0
	BICICICLE	1
	BICIRCLE	0
	BICYRCLE	2
	BICYKLE	6
	BICYELE	50
	BICCYLE	1
	BICYLCE	42
	BICYC!E	6
	BICYCE	13
	BIICYCLE	1
	BIYCYCLE	4
	BIVCYCLE	45
	BIECYCLE	4
	BICIYCLE	1
	BICLYCLE	3
	BICYCILE	4
	BICSYCLE	1
	BICYLCLE	14
	BICYCVLE	13
	BICYCELE	6
	BICYCLIES	4
	BICYCLYE	1
	BICYCLLE	2
	BICYCLEE	1
	BICYCI	2
	BICYDE	42
	BICYCCLETTE	1
	BICYCLEELETTE	1
	BICICLE CLETTE	1
	BICYYCLETTE	1
	EBICYCLE	4

Table 7. Misspellings of bicycle in applicant and titles.

Misspellings in the applicant	WO2015/005943	SLIPSTREAM BYCYCLES
	US20070010376	TAIWAN BICYLCE INDUSTRY R & D CENTER
	EP2103512	CANNONDALE BICYLE
Misspellings in the title	EP0825101	Electric bicyle
	CA3053537	Biycycle seat post travel adjustment assembly

Table 8. Keywords (correct and with misspellings) used to search patents with the applicant BOSCH and number of patents retrieved for each of them.

Correct keyword	(BOSCH AND (ROBERT 1W BOSCH)) /PA/OPA (124538 patents)
Keywords with misspellings	BOSH (6), BOSCHE (5), BOSGH (2), BOCH (2), BOECH (2), BOACH (1), BOSTH (1), BOBCH (1), BOEOH (1), BOCSH (1)

Considering that Bosch is a name of only 5 letters, 10 different misspellings were found.

The most famous companies are often subject to continuous monitoring systems by competitors. Just think of Apple's products and how much resonance in the newspapers there is as soon as they find out from patents what it is working on. The secrecy period for a patent lasts a maximum of 18 months. To bypass this alerting system that is actually set up on the applicant's name, it is enough to misspell the applicant's name on the patent application and hope no one notices. Given the number of instances of errors on applicants, it is safe to assume that they are not all oversights resulting from distraction.

The number of variants of the same applicant increases when the syllables of his name increase (e.g., for SIEMENS we found applicants called SIMENS, SIEMEN, SIIEMENS, SIEMEMS, SIEMMENS, SIEIMENS, STIEMENS, SIEDMENS, SIEMIENS, SIEMEENS, SIEMYENS, SIEMUENS, SIEMEINS, SIEMENTS, SIEMENYS, SIE3MENS, SIEM1ENS, SIEME3NS, SIEMENSA, SIEMENSAG, although it is not excluded that some of these variants may lead to real companies). The most complex situations occur with long company names in combination in the form of multiword, e.g., Hewlett Packard. Other disruptive factors, which can lead to search errors, are the applicant's acronyms, e.g., "HP Inc." which may themselves contain misspellings.

5.4. CO_2 Search and CARBON DIOXIDE Search

Table 9 reports the results, in terms of patents, obtained using the queries with the correct keyword (i.e., "(CO_2)/ETI/ECLM/DESC" and "(CARBON 0W DIOXIDE)/ETI/ECLM/DESC") and with the keywords with the misspellings.

In both cases, although percentage-wise the number of patent typos is small compared to the total, the absolute number is definitely significant. A search that does not take into account these variant spellings may lead to entirely erroneous considerations and results.

5.5. SUPERCRITICAL Fluid Search

Table 10 reports the results, in terms of patents, obtained using the queries with the correct keyword (i.e., "SUPERCRITICAL/ETI/ECLM/DESC") and with the keywords with the misspellings.

With more complicated multiword, the number of documents reporting misspellings increases considerably, in the case of supercritical we are around 8% and we have not considered them all. Such high percentages can undermine all kinds of research.

5.6. SERGEI ALEXANDROVICH IKOVENKO Search

Sergei Alexandrovich Ikovenko claims on his website that he is the inventor of 104 patents. In this case, the search for the inventor using his name written in Cyrillic, i.e., "(ЯКОВЕНКО AND (СЕРГЕЙ OR АЛЕКСАНДРОВИЧ))/IN/OIN/INH/INV" in Orbit database (dove ЯКОВЕНКО = Ikovenko, СЕРГЕЙ = Sergei, АЛЕКСАНДРОВИЧ = Alexandrovich) provided only three results. Even the search for the English name "(IKOVENKO AND (SERGEI OR ALEXANDROVICH))/IN/OIN/INH/INV" provided only three results.

Table 9. Keywords (correct and with misspellings) used to search patents about CO_2 and CARBON DIOXIDE.

Keywords		Results (N° of Patents)		Results (N° of Patents)
Correct keyword	CO_2	536,459		
Keywords with misspellings	CO 2; CO-2; CO.2; CO_2-; CO_2~; CO_2; CO.2; CO_2; CO–2; CO:2	831,022		
	C02 (Zero instead of O)	76,624		
	C0 2; C0-2; C0.2; C02-; C02~; $C0_2$; C0.2; C0_2; C0–2; C0:2	118,156		
	Cθ2 (Teta instead of O)	2288		
	Cθ 2; Cθ-2; Cθ.2; Cθ2-; Cθ2~; $Cθ_2$; Cθ.2; Cθ_2; Cθ–2; Cθ:2	3982		
	CO sub 2; CO.sub.2; CO.sub_2; CO sub_2	20,482		
	C0 sub 2; C0.sub.2; C0.sub_2; C0 sub_2	25		
Correct keyword	CARBON DIOXIDE	1,164,394		
Keywords with misspellings providing more patents (>1000 each)	DIOXIDE CARBON	42,424		
	CARBON OXIDE	31,867		
	CARBONDIOXIDE	4279		
	CARBON DIOXID	2888		
	CARBON DI OXIDE	1990		
Other keywords with misspellings	CARBON BIOXIDE	200	CARBON DIOX7D	2
	CARBON BI OXIDE	12	CAROBN DIOXIDE	19
	CARBON DOUBLE OXIDE	2	CARBBN DIOXIDE	27
	CARBON TWO OXIDE	2	CACBON DIOXIDE	6
	CARBN DIOXIDE	25	CARTON DIOXIDE	120
	CARBONE DIOXIDE	148	CANBON DIOXIDE	5
	CARBONIC DIOXIDE	263	CAHON DIOXIDE	3
	CARBONIC DI OXIDE	6	CARBEN DIOXIDE	38
	CARBON DIOXYDE	146	CATBON DIOXIDE	48
	CARBON DIOXDE	178	CARLBON DIOXIDE	34
	CARBON DEOXIDE	34	CARHON DIOXIDE	83
	CARBONS DIOXIDE	50	CARBAN DIOXIDE	28
	CARBON DI OXID	60	CARLION DIOXIDE	10
	CARBONDIOXID	107	CAIRBON DIOXIDE	20
	CARON DIOXIDE	44	CARBON DIOXCDE	1
	CARTOON DIOXIDE	31	CARBOIN DIOXIDE	9
	CARBON DIOXODE	58	CARBON DIOXFDE	5
	CARHON DI OXID	1	CARIBON DIOXIDE	20
	CARHON DIOXID	83	CAFBON DIOXIDE	15
	CARDON DIOXIDE	41	CARBDN DIOXIDE	7
	CARBLON DIOXIDE	32	CANON DIOXIDE	5
	CAROON DIOXIDE	40	CAMBON DIOXIDE	4
	CABON DIOXIDE	132	CARDAN DIOXIDE	1
	CARBON DIOXJD	3		

To figure out where all the patents ended up, it was necessary to apply transliteration rules from Cyrillic to the Latin alphabet. For example, Ikovenko's I can appear under the following combinations (I = J = Y = Yi = Yy = Iy = Ii = Ij), the Я on the other hand (Я= Ya = Yia = Ja = Ia = Iya = Jja = Iia = Yya = Yja = Jia = Jya = Ija) and finally ko is also Ko = co = cho. The complete list of substitutions is shown in Figure 1.

Only the query by inventor, formed by the intersection (AND) between the union of the set of surnames and the union of the set of first names and the set of second names, it provided all 104 patents. Where the set of surnames includes the surname in Cyrillic and the surname in English with all the misspellings coming from transliterations.

In particular, the only four terms "IKOVENKO, YAKOVENKO, JAKOVENKO e YACOVENKO", joined together by the "OR" logical operator provide 1729 patents, equal to 99% of all the patents obtainable with the query made up of all the surnames.

Figure 2 graphically represents the query used to search for the inventor.

Table 10. Keywords (correct and with misspellings) used to search patents about SUPERCRITICAL fluids.

Keywords		**Results (N° of Patents)**		**Results (N° of Patents)**
Correct keyword	SUPERCRITICAL	158,899		
Keywords with misspellings providing more patents (>100 each)	SUPER CRITICAL	10,627		
	SUPERCRITICALLY	1213		
	SUPERCRITICALITY	364		
	SUPERIOR CRITICAL	120		
	SUPER CRITICALLY	108		
Other keywords with misspellings	SUPERCRITICALS	12	SUPERCCRITICAL	1
	SUPER CRITICALS	1	SUPERECRITICAL	2
	SUPERCRITIC	90	SUPERHEAT CRITICAL	6
	SUPER CRITIC	7	SUPERHEATER CRITICAL	4
	SUPERCRITICALNESS	1	SUPERICRITICAL	2
	SUPERCRITICALY	2	SUPERIORCRITICAL	1
	SUPERCRITICF	1	SUPERLCRITICAL	1
	SUPERCRITICISM	1	SUPERRCRITICAL	2
	SUPER CRITICISM	2	SUPERS CRITICAL	1
	SUPERCRITICALL	5	SUPERSCRITIC	1
	SUPERCRITICA	51	SUPERSCRITICAL	3
	SUPER CRITICA	9	SUPERSUPERCRITICAL	77
	SUPERCRYTICAL	5	SUPERTCRITICAL	2
	SUPER CRITICALITY	28	SUPERTRICRITICAL	9
	SUPER CRITCAL	4	SUPPER CRITICAL	15
	SUPERCRITCAL	60	SUPPERCRITICAL	3
	SUPERCRITICALIZATION	3	SUPRA CRITICAL	23
	SUPERCRYTICALLY	2	SUPRA CRITICALLY	3
	SUPERCRTICAL	22	SUPRACRITIC	1
	SUPERC CRITICAL	1	SUPRACRITICAL	25
	SUPRACRITICALLY	4		

Transliteration errors are the most complex to find. Searching by an inventor in general is a search that never guarantees adequate recall and precision. It is no coincidence that to help this kind of search, avoiding problems dealing with homonyms, misspellings, and inversion of the first name with the last name, activities are being promoted especially in scientific publications and academics to file authors by a unique identifying code, e.g., Orcid [31].

5.7. Final Considerations

Figure 3 graphically represents the comparison between the retrieved patents with and without misspellings in each considered patent search, i.e., description/claims/title fields, by using acronym, single words, multi-words about a technology (i.e., BRAYTON CYCLE, BICYCLE) or a material (i.e., CO_2, SUPERCYTICAL fluid, CARBON DIOXIDE), in the applicant field (i.e., BOSCH) and in the inventor field (i.e., SERGEI ALEXANDROVICH IKOVENKO).

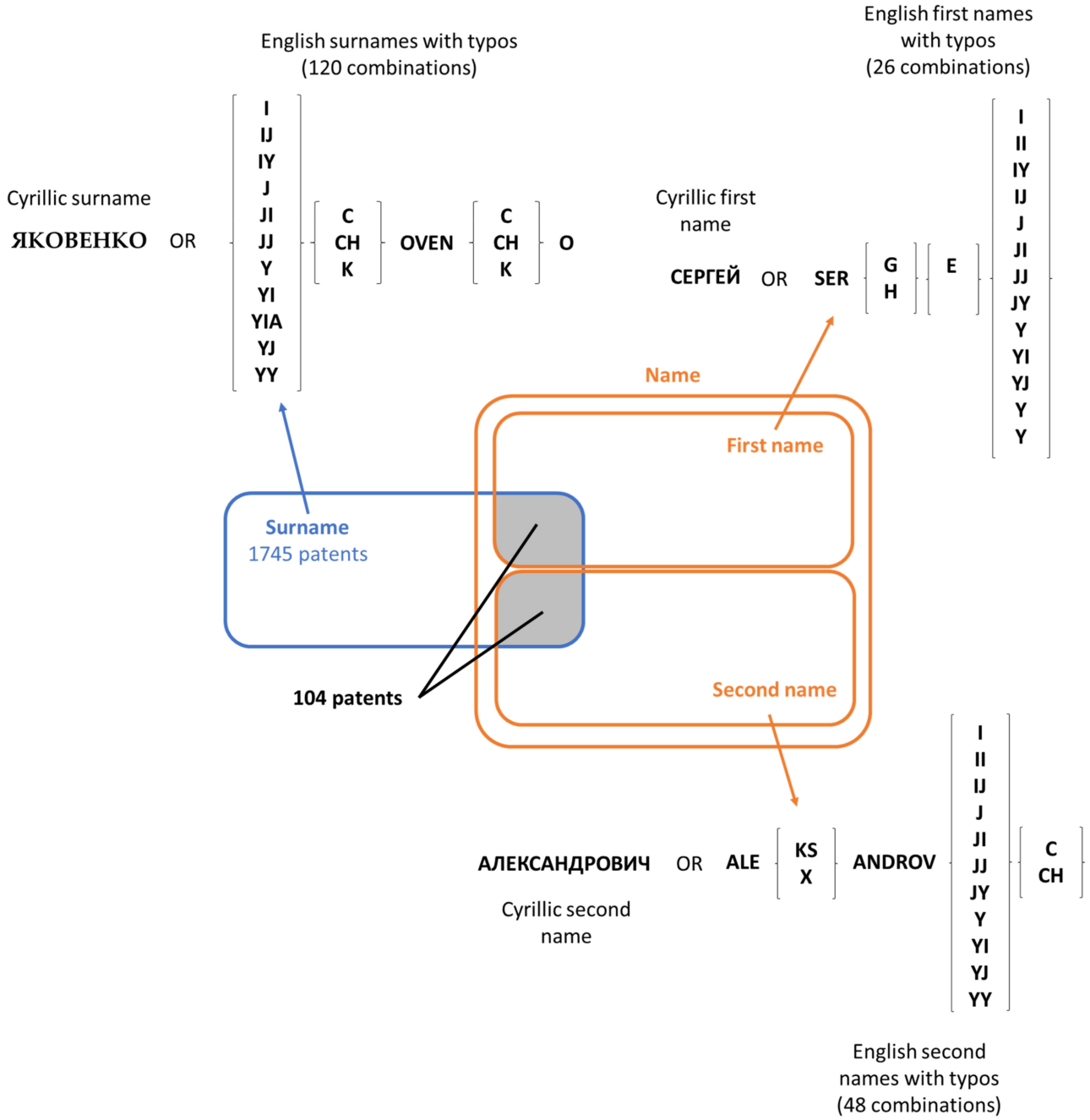
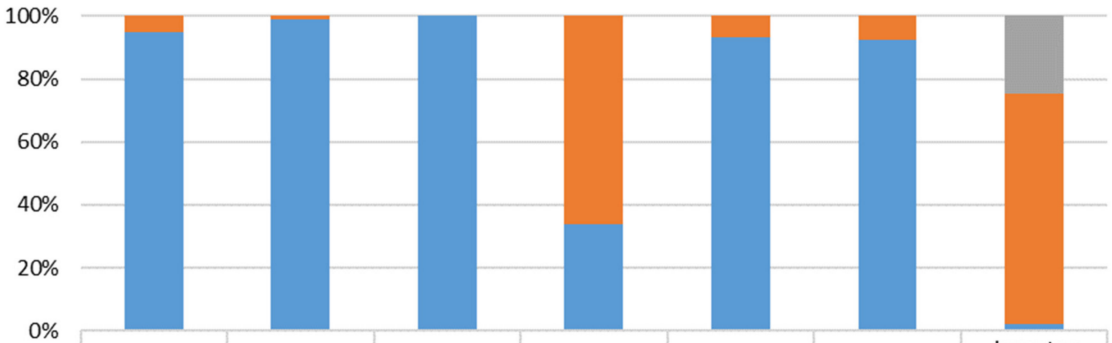


Figure 2. Graphical representation of the possible combinations of misspellings in the search for the name of an inventor.

The cases presented, although limited in number, are in our opinion useful for extrapolating some indications that are of general significance. In all the proposed searches, with rare exceptions such as searches by inventor or acronyms and chemical formulas, typos affect very small percentages. These percentages although small cannot always be considered negligible, since lost patents can still affect the final opinion for those searches such as prior art or freedom to operate.

In addition, searches for misspellings in multiwords, even where the result percentages are small, represent in absolute numbers an important patent pool. Disregarding these document pools can undermine the reliability of technical analyses such as technological survey, forecasting, patent intelligence, etc.

Searching by inventor, especially if the name is translated from languages other than Latin alphabet follows ad hoc rules that cannot be disregarded to avoid completely misplacing the search outcome.



	Technology search (multiwords): BRAYTON CYCLE	Technology search (single words): BICYCLE	Applicant search: BOSCH	Material search (acronym): CO2	Material search (multiwords): CARBON DIOXIDE	Technology search (single words): SUPERCRITICAL (fluid)	Inventor search: SERGEY ALEXANDROVICH IKOVENHO
Patent containing typos	6%	1%	0%	196%	7%	8%	3367%
Retrieved patents with misspellings	300	2,335	22	1,052,579	85,656	12,925	101
Retrieved patents without misspellings	5,395	188,699	124,538	536,459	1,164,394	158,899	3

Figure 3. Graphical representation of the retrieved patents with and without misspellings in the considered patent searches.

6. Discussions

The obtained results, presented in Section 4, provided evidence to confirm HP. 1 (i.e., misspellings can be present in all the parts of the text of a patent, e.g., description vs applicant or inventor name, etc.) since it was easy to demonstrate with absolutely generic examples of how no part of the patent text is free of misspellings.

Meanwhile, the confirmation of HP. 2 (i.e., different types of misspellings, contained in the same or different parts of a patent text, can influence the information retrieval from patents more or less severely) is obtained by analyzing and discussing in detail the different types of misspellings that have been obtained.

In fact, not all typos act in the same way. As it is easy to guess from the choice of the proposed cases, those that impact the most are precisely those with Levenshtein distance of one, that is, those that deviate little from the correct word. Among these misspellings, it is the repetitive ones that make the preponderant part, often due to a phonetic error or not fully knowing the rules of compound construction.

A major limitation of this approach is that it is not always possible to search directly for those misspellings that include the same truncation symbols used to search for it. As much as we can counteract the proliferation of errors, no technique will ever guarantee a 100 percent safe result, however, awareness coupled with the combination of good tools can be of great help. Fortunately, the keyword searched for is never mentioned only once in the entire document, and if there is no intentional misspelling on purpose; there is also a good chance that it will be spelled right at least once, thus limiting the damaging effect of misspellings on the final result.

7. Conclusions

This study systematically analyzed the types of misspellings present in the text of a patent to show how their presence can influence the patent search and therefore the search for information in patents.

Beyond the limitations relating to the type and quantity of errors considered (based only on the minimum leverage distance and excluding special characters) and the number of case studies considered, this study confirmed both starting hypotheses:

- Misspellings can be present in all the constituent parts of a patent, i.e., description, title, claims, applicant, and inventor.
- Different misspellings, voluntary or accidental, single, or repeated, affect the patent search differently. The biggest problem is that the longer or more complex the words are, the greater the number of possible misspellings. Some misspellings are easily identifiable and allow to identify the patent sought, others not, or worse, they lead to the identification of wrong and misleading patents.

In conclusion, this study, therefore, warns about the role of misspellings in patent search and highlights their negative impact after having duly classified the misspellings and isolated the most critical cases. This result can be the first knowledge base to identify the best tools to identify and bypass the problems of misspellings. In this article, we recommend a combination of complementary tools, tools implementing word and contextual embedding models, and rule-based tools specifically designed on the linguistic rules most likely to cause errors.

The proposed case studies do not claim to be exhaustive but have been chosen to represent the main error scenarios. It is certainly possible to increase the number of case studies considered, as well as to expand the set of documentary sources outside of patent research.

The research produced so far has been suggested for patent analysis because it is the research on which it is easier to verify the impact and because it is a widely used technical database, with free access for everyone, e.g., Espacenet. In reality, it is easily demonstrable that the presence of misspellings afflicts every documentary source. Scientific articles are not exempt, even though the editors produce macros with automatic error recognition and that each document undergoes several review steps. The authors conducted similar investigations both on the text of scientific journals and on funded projects such as the database of Horizon 2020 financed projects of the European Community (https://cordis.europa.eu/) (accessed on 28 July 2022) finding the same dynamics described for patents.

These points are the planned future developments of this study, in order to take a more realistic picture of misspellings in patents and beyond.

Author Contributions: Conceptualization, D.R.; methodology, D.R.; software, D.R. and A.P.; validation, S.A.; formal analysis, D.R.; investigation, D.R. and S.A.; resources, D.R.; data curation, S.A. and A.P.; writing—original draft preparation, D.R. and C.S.; writing—review and editing, D.R. and C.S.; visualization, C.S.; supervision, D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The tool used to do the tests, in this case, was kindly provided by TRIX srl, a company specializing in the creation of search engines.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Stein, B.; Hoppe, D.; Gollub, T. The impact of spelling errors on patent search. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, 23–37 April 2012; pp. 570–579.
2. Adams, S. The Text, the Full Text and nothing but the Text: Part 1–Standards for creating Textual Information in Patent Documents and General Search Implications. *World Pat. Inf.* **2010**, *32*, 22–29. [CrossRef]
3. Moon, J.; Burstein, F. Ontology-based spelling correction for searching medical information. In *Semantic Web Technologies and E-business: Toward the Integrated Virtual Organization and Business Process Automation*; IGI Global: Hershey, PA, USA, 2007; pp. 384–404.
4. Bhole, A.; Udupa, R. On correcting misspelled queries in email search. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
5. Huang, B. WNSpell: A WordNet-based spell corrector. In Proceedings of the 8th Global WordNet Conference (GWC), Bucharest, Romania, 27–30 January 2016; pp. 136–143.
6. Hossain, M.M.; Labib, M.F.; Rifat, A.S.; Das, A.K.; Mukta, M. Auto-correction of english to bengali transliteration system using levenshtein distance. In *2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, 28–30 June 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
7. Suzuki, H.; Gao, J. A comparative study on language model adaptation using new evaluation metrics. In Proceedings of the EMNLP, Vancouver, BC, Canada, 6–8 October 2005.
8. Tokunaga, H.; Okanohara, D.; Mori, S. Discriminative method for Japanese kana-kanji input method. In Proceedings of the Workshop on Advances in Text Input Methods (WTIM 2011), Chiang Mai, Thailand, 25 July 2011.
9. Suzuki, H.; Gao, J. A unified approach to transliteration-based text input with online spelling correction. In Proceedings of the EMNLP, Stroudsburg, PA, USA, 12–14 July 2012.
10. Lee, J.-H.; Minho, K.; Hyuk-Chul, K. Deep learning-based context-sensitive spelling typing error correction. In *IEEE Access*; IEEE: Piscataway, NJ, USA, 2020; Volume 8, pp. 152565–152578.
11. Cohen, W.; Ravikumar, P.; Fienberg, S. A comparison of string metrics for matching names and records. In Proceedings of the Kdd Workshop on Data Cleaning and Object Consolidation, Washington, DC, USA, 7–12 August 2003; Volume 3, pp. 73–78.
12. Knuth, D.E. *The Art of Computer Programming, Volume I: Fundamental Algorithms*, 3rd ed.; Addison-Wesley: Boston, MA, USA, 1997.
13. Stein, B.; Curatolo, D. Phonetic Spelling and Heuristic Search. In *Frontiers in Artificial Intelligence and Applications*; IOS Press: Amsterdam, The Netherlands, 2006; Volume 141, p. 829.
14. Lee, J.-H.; Minho, K.; Hyuk-Chul, K. Improved statistical language model for context-sensitive spelling error candidates. *J. Korea Multimed. Soc.* **2017**, *20*, 371–381. [CrossRef]
15. Lewellen, M. Neural network recognition of spelling errors. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, QC, Canada, 10–14 August 1998; Volume 2, pp. 1490–1492.
16. Chrupała, G. Normalizing tweets with edit scripts and recurrent neural embeddings. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; Volume 2, pp. 680–686.
17. Al Azawi, M.; Ul Hasan, A.; Liwicki, M.; Breuel, T.M. Character-level alignment using WFST and LSTM for post-processing in multi-script recognition systems-A comparative study. In Proceedings of the International Conference Image Analysis and Recognition, Loulé, Portugal, 22–24 October 2014; Springer: Cham, Switzerland, 2014; pp. 379–386.
18. Liu, F.; Weng, F.; Wang, B.; Liu, Y. Insertion, deletion, or substitution? Normalizing text messages without pre-categorization nor supervision. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 71–76.
19. Contractor, D.; Faruquie, T.A.; Subramaniam, L.V. Unsupervised cleansing of noisy text. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Beijing, China, 23–27 August 2010; Association for Computational Linguistics: Stroudsburg, PA, USA, 2010; pp. 189–196.
20. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 26–28 October 2014.
21. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of tricks for efficient text classification. *arXiv* **2016**, arXiv:1607.01759. [CrossRef]
22. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]
23. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv* **2019**, arXiv:1910.13461.[CrossRef]
24. Gardner, M.; Grus, J.; Neumann, M.; Tafjord, O.; Dasigi, P.; Liu, N.; Peters, M.; Schmitz, M.; Zettlemoyer, L. Allennlp: A deep semantic natural language processing platform. *arXiv* **2018**, arXiv:1803.07640.[CrossRef]
25. Floridi, L.; Chiriatti, M. GPT-3: Its nature, scope, limits, and conse-quences. *Minds Mach.* **2020**, *30*, 681–694. [CrossRef]
26. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805. [CrossRef]

27. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. In Proceedings of the 33rd Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 August 2019.
28. Russo, D.; Carrara, P.; Facoetti, G. Technical problem identification for supervised state of the art. *IFAC-PapersOnLine* **2018**, *51*, 1341–1346. [CrossRef]
29. Montecchi, T.; Russo, D.; Liu, Y. Searching in Cooperative Patent Classification: Comparison between keyword and concept-based search. *Adv. Eng. Inform.* **2013**, *27*, 335–345. [CrossRef]
30. Kuznetsov, A.; Urdiales, H. Spelling correction with denoising transformer. *arXiv* **2021**, arXiv:2105.05977. [CrossRef]
31. Haak, L.L.; Fenner, M.; Paglione, L.; Pentz, E.; Ratner, H. ORCID: A system to uniquely identify researchers. *Learn. Publ.* **2012**, *25*, 259–264. [CrossRef]