International Standard

ISO 5078

First edition
2025-02

Management of terminology resources — Terminology extraction

Gestion des ressources terminologiques — Extraction de terminologie

Reference number
ISO 5078:2025(en)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 37, *Language and terminology*, Subcommittee SC 3, *Management of terminology resources*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

Over the past decades, extracting relevant designations, mostly terms (i.e. linguistic designations), from text corpora has become an increasingly important task carried out in a wide variety of different fields. Terminology extraction, which goes beyond mere extraction of terms, is undertaken by a range of specialists including language professionals in general, and terminologists in particular, as well as ontology engineers, and both information and data scientists. Terminology extraction also serves several purposes that go beyond the compilation of glossaries or the population of terminology databases, including the identification of concepts and of concept relations for building ontologies.

The widespread use of terminology extraction tools in terminology management, as well as in other fields such as information retrieval, stands in stark contrast to the rarity of individual documents that provide definitions, requirements or best practices.

However, although terminology extraction tools save time, money and effort in terminology management, their output becomes even more relevant when it is assessed and validated, using both qualitative and quantitative approaches and criteria for selecting entities such as relevant terms, definitions and concept relations. This extracted and then validated terminological data supports the building of high-quality terminology resources and, thus, terminology management.

This document covers the following aspects that form the core of terminology extraction methods and practices in general:

— compilation of text corpora (general principles and types of text corpora);

— methods and criteria employed by mainstream terminology extraction tools (statistical, linguistic, hybrid and neural);

— criteria for selecting terms (filtering candidate term lists and assessment of term eligibility);

— tool characteristics.

By objectively specifying these aspects, this document provides a reference framework for improving the performance of terminology extraction tools and optimizing the use of their output.

Management of terminology resources — Terminology extraction

1 Scope

This document specifies methods for extracting candidate terms from text corpora and gives guidance on selecting relevant designations, definitions, concept relations and other terminology-related information.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 704, *Terminology work — Principles and methods*

ISO 1087, *Terminology work and terminology science — Vocabulary*

ISO 16642, *Management of terminology resources — Terminological markup framework*

ISO 26162-1, *Management of terminology resources — Terminology databases — Part 1: Design*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
annotation
process of adding *metadata* (3.10) to segments of language data

[SOURCE: ISO 24617-1:2012, 3.2, modified — “information” replaced by “metadata”; “or that information itself” deleted.]

3.2
bitext
collection of *texts* (3.24) in two languages that can be considered translations of each other and that are segmented and aligned

Note 1 to entry: Bitexts play a key role in training, evaluating and improving localization technologies, such as translation memories, terminology management tools or machine translation engines.

3.3
candidate term
term candidate
provisional term
string of *characters* (3.5) that has been collected by means of *term extraction* (3.20) but has not yet been selected as a relevant *term* (3.19) to be considered for inclusion in a *terminological data* (3.22) collection

[SOURCE: ISO 12616-1:2021, 3.18, modified — "text element to be documented in the" replaced by "term to be considered for inclusion in a".]

3.4
candidate terminological data
string of *characters* (3.5) that has been collected by means of *terminology extraction* (3.23) but has not yet been selected as relevant *terminological data* (3.22)

3.5
character
unit of textual information represented by one or more bytes

EXAMPLE Single letter, numeral, punctuation mark, diacritic, symbol, ideograph, space.

[SOURCE: ISO/IEC 14840:1996, 4.10, modified — "textual" added to the definition; example added.]

3.6
collocation
lexically or pragmatically constrained recurrent cooccurrence of at least two *lexical units* (3.8) which are in a direct syntactic relation with each other

EXAMPLE "Commit a crime" instead of "do a crime".

3.7
keyness
quantity proportional to the frequency of a *lexical unit* (3.8) in a subject-field-specific *text corpus* (3.25), relative to a *reference corpus* (3.15)

3.8
lexical unit
meaningful element in the *lexicon* (3.9) of a language

3.9
lexicon
complete set of meaningful elements in a language

3.10
metadata
data that defines and describes other data

[SOURCE: ISO 24531:2013, 4.32]

3.11
n-gram
sequence of *n* adjacent *tokens* (3.27)

Note 1 to entry: Frequently adjacent tokens can be an indicator for *termhood* (3.21).

Note 2 to entry: The number of adjacent tokens (*n*) is usually 2, 3 or 4.

3.12
noise
non-relevant search results

Note 1 to entry: In *terminology extraction* (3.23), "noise" means non-relevant data in the extraction output.

3.13
precision
ratio of relevant search results to all search results

Note 1 to entry: In *terminology extraction* (3.23), "precision" means the ratio of relevant *candidate terms* (3.3) retrieved to the total of candidate terms retrieved.

Note 2 to entry: Precision and *recall* (3.14) generally have an inverse relationship; when one increases, the other tends to decrease.

3.14
recall
ratio of relevant search results to all relevant items in a set that have been or should have been found from a search query

Note 1 to entry: In *terminology extraction* (3.23), "recall" means the relevant *candidate terms* (3.3) in a *text corpus* (3.25).

Note 2 to entry: Recall and *precision* (3.13) generally have an inverse relationship; when one increases, the other tends to decrease.

3.15
reference corpus
text corpus (3.25) to which a given text corpus for *terminology extraction* (3.23) is compared

3.16
relevance
quality of being a successful search result in relation to the search query

3.17
silence
set of relevant search results that have not been found from a search query

Note 1 to entry: In *terminology extraction* (3.23), "silence" means the set of valid *candidate terms* (3.3) that are missing in the extraction results.

3.18
stop word
word that is not taken into account as a *candidate term* (3.3)

Note 1 to entry: Typical stop words are function words (e.g. prepositions, articles), brand names and non-special language words to the specific subject field.

3.19
term
designation that represents a general concept by linguistic means

EXAMPLE "laser printer", "planet", "pacemaker", "chemical compound", "¾ time", "Influenza A virus", "oil painting".

Note 1 to entry: Terms can be partly or wholly verbal.

[SOURCE: ISO 1087:2019, 3.4.2]

3.20
term extraction
identification and excerption of *candidate terms* (3.3)

Note 1 to entry: *Terms* (3.19) can include all types of designations, including appellations, proper names and symbols.

3.21
termhood
degree to which a *lexical unit* (3.8) is recognized as a *term* (3.19)

EXAMPLE "Mouse" has stronger termhood in computer applications and weaker termhood in general language.

Note 1 to entry: Termhood applies to both simple terms (consisting of a single word) and complex terms (consisting of more than one word or lexical unit), and to other designations, such as proper names and appellations, as well as formulas and symbols.

[SOURCE: ISO 26162-3:2023, 3.13, modified — Example revised.]

3.22
terminological data
data related to concepts and their designations

Note 1 to entry: Common terminological data include designations, definitions, contexts, notes to entry, grammatical labels, subject labels, language identifiers, country identifiers, and source identifiers.

[SOURCE: ISO 1087:2019, 3.6.1]

3.23
terminology extraction
identification and excerption of *candidate terminological data* (3.4)

3.24
text
content in written form

[SOURCE: ISO 20539:2023, 3.3.1]

3.25
text corpus
collection of natural language data

[SOURCE: ISO 1087:2019, 3.6.4, modified — Admitted term and Note 1 to entry deleted.]

3.26
TF-IDF
term frequency — inverse document frequency
statistical value intended to reflect how important a *lexical unit* (3.8) is to a document in a *text corpus* (3.25)

3.27
token
individual occurrence of a *type* (3.29) in a *text corpus* (3.25)

3.28
tokenization
conversion of *text* (3.24) into *tokens* (3.27)

3.29
type
unique sequence of *characters* (3.5) in a *text corpus* (3.25)

Note 1 to entry: The number of types is different from the number of occurrences (*tokens* (3.27)).

Note 2 to entry: While the number of tokens in a text corpus refers to the total number of occurrences, the number of types refers to the total number of unique occurrences.

3.30
unithood
degree to which a given sequence of words has sufficient collocational strength to form a stable *lexical unit* (3.8)

EXAMPLE "Art deco table" has stronger unithood than "modern table".

Note 1 to entry: Because unithood derives from the collocational relationship of words making up a given string, it only applies to multi-word *terms* (3.19).

[SOURCE: ISO 26162-3:2023, 3.15]

3.31
validated term
candidate term (3.3) which meets specified criteria

3.32
validated terminological data
candidate terminological data (3.4) which meets specified criteria

3.33
vector
quantity having direction as well as magnitude

[SOURCE: ISO 19123-1:2023, 3.1.51, modified — Note 1 to entry deleted.]

3.34
vector space model
statistical model for representing text information as a *vector* (3.33) of identifiers

Note 1 to entry: Vector space models can be used for information retrieval (IR), natural language processing (NLP) or text mining tasks in order to identify whether *texts* (3.24) are similar in meaning.

[SOURCE: Reference [15], modified — "for Information Retrieval, NLP, Text Mining" moved from the definition to Note 1 to entry; "as a vector of identifiers" added to the definition; Note 1 to entry added.]

4 Principles and methods

4.1 General

Terminology extraction requires a deep understanding of terminology theory and terminology work. In this sense, and to achieve high-quality results, the following shall be used:

— established terms and definitions as specified in ISO 1087;

— principles and methods as specified in ISO 704;

— data-modelling criteria as specified in ISO 16642;

— terminology database design principles as specified in ISO 26162-1.

There are various types of text corpora. Selection of corpus type and texts to be included is usually influenced by factors such as project goal, scope and deadlines.

4.2 Text corpora and terminology extraction

Organizations usually produce textual material relating to their industry, activity and the field in which they operate. These kinds of texts include, for example, marketing materials, product documentation, internal memos and bilingual translation memories. Such textual material can contribute to an organization-wide text corpus that forms the basis for terminology extraction.

The usefulness of candidate terminological data extracted from such a text corpus depends on the context and aim of the terminology extraction project as well as on the depth or breadth of the subject-field coverage provided by the text corpus.

4.3 Compilation of text corpora

4.3.1 Text corpora used for terminology extraction

Terminology extraction begins with the collection of a text corpus, according to the objectives of the project. There are differing kinds of text corpora, specifically:

— a monolingual corpus, consisting of texts taken from the same language;

— a bilingual corpus, consisting of texts taken from two languages;

— a multilingual corpus, consisting of texts taken from more than two languages;

— a parallel corpus, consisting of texts taken from one language aligned with their translations into one or more other languages;

EXAMPLE A set of annual reports in English aligned segment by segment with the same annual reports translated into Tagalog.

— a comparable corpus, consisting of one or more sets of texts meeting certain criteria, matched against a set of texts meeting the same criteria in one or more languages.

NOTE Thus, the texts are not translations of each other, but they are similar in respects other than language, which can be, for example, subject field or text type. A comparable corpus can be created, for example, from original articles on steel extrusion processes in French and another set of original articles on steel extrusion processes in German.

When creating a text corpus, texts should be selected depending on the goal or purpose of the extraction task by defining and/or selecting criteria the texts must meet to be included in the text corpus. For example, if the goal is to extract rare words in Elizabethan English, a text that is a German computer manual should not be included in the text corpus because it does not meet two criteria essential and appropriate to the goal: the time frame criterion (texts written between 1558 and 1603) and the language criterion (in English). Or if the goal is to translate the user interface of a software program into another language, creating a text corpus out of the textual graphical user interface (GUI) elements can be useful for extracting terms. As a next step, language professionals can find equivalents for these terms to be used in the translation.

4.3.2 Criteria for selecting texts for a text corpus

Terminology extraction can have a number of different objectives, all of which influence the development of the text corpus. The following list includes frequent criteria that can be considered when selecting texts for a text corpus. This list is based on scenarios that seek to extract and use terminology for the same purpose associated with the texts in the corpus:

— Content source: Usually, original content is preferable to derived works or content generated by artificial intelligence applications.

— Language originality: In monolingual extraction scenarios, texts for a given language which have originally been written in that language are preferable to translated texts. Translated texts can be used, particularly where original texts are not available (e.g. in cases of languages where very little has been written in a particular subject field). In bilingual extraction scenarios, however, original texts are usually aligned with their corresponding translations.

— Locale: When seeking to extract terminology particular to a locale (language plus region), including texts from only that locale provides the best results.

EXAMPLE 1 Canadian French texts for Canadian French terminology, Swiss French texts for Swiss French terminology.

— Scope: When seeking to extract terminology applicable to a particular part of a subject field, selecting texts destined to be used in that subject field part yields the best results.

EXAMPLE 2 When seeking terms relating to oncology, selecting oncology-related texts from a medicine text corpus over dermatology-related ones will generate more appropriate results.

— Intended audience: Selecting texts to extract terminology that fits the expectations and needs of the intended users of the extracted terminology will yield better results.

EXAMPLE 3 If the specifications dictate plain language for a general, non-specialist audience, the selection of texts for the corpus would differ from texts that would be chosen if the audience were skilled professionals.

— Time frame: When seeking to extract terminology of a given time period, selecting texts that have been created in that period will provide the most relevant results.

— Representativeness: Selecting texts that are relevant to the community of experts of a specific subject field leads to more relevant results. Texts that are representative of the subject field(s) for which terminology is to be extracted are preferable to texts that mention the subject field(s) tangentially.

— Authority: Peer-reviewed documents published by a recognized authority are preferable to other documents. Texts written by subject-field experts are usually preferable to texts created by non-subject-field experts. Depending on the objectives, however, it can be useful to extend the scope to include other authors.

— Language register: Selecting texts according to the purpose of the particular communication situation (e.g. formal language used in laws versus informal language used in text messages) will result in more relevant data being extracted.

— Document type: When seeking to extract terminology specific to a particular type of document, limiting the selection to documents that belong to that document type (e.g. web pages, manuals, reference books) will yield more relevant results.

— File size: Choosing documents that contain a reasonable amount of terminology with regard to the type of terminology project is preferable. One short file often does not contain enough terminology to be extracted if a tool only extracts candidate terms that appear at least twice in a text corpus. It is useful to keep in mind how quickly the terminology extractor can process large files as well as how powerful the computer running the extraction software is. If the files are too large and the computer or tool too slow, terminology processing can take more time than has been allotted to a terminology extraction task.

— File format: Digitized documents are preferable to scanned documents. It saves time to select documents in formats that can be processed by the terminology extraction tool being used (or that can quickly and easily be converted).

As stated, the aforementioned criteria hypothesize extraction scenarios that aim at reusing the extracted terminology in content production situations that are similar to the texts in the text corpus. However, sometimes it can be necessary to adjust criteria (e.g. if no digitized documents are available, if the goal is to describe how terminology evolved over time, if the extracted terminology is used to depict the inappropriate use of terminology).

When building a text corpus, the formats in which documents are available and the formats the terminology extraction tool can handle limit the amount of text that can be included in a text corpus. While some tools have embedded conversion features, for others, file conversion tools help widen the set of possible documents that can be used by converting files from formats the extraction tool cannot handle into a format that can be handled by the tool.

In summary, the choice of texts and text types to include will depend on the goals of the terminology extraction project and the criteria and parameters chosen.

4.3.3 Considerations for text corpus creation

To extract specific terms, creating a corpus of recent and authoritative texts can be useful. For example, to extract company-specific terms, it can be useful to build a corpus of recent, authoritative texts from the company's intranet and website.

If planning to create a bilingual termbase of those organization-specific terms, then creating a corpus of all original texts for which there is a translation as well as their translations should be considered.

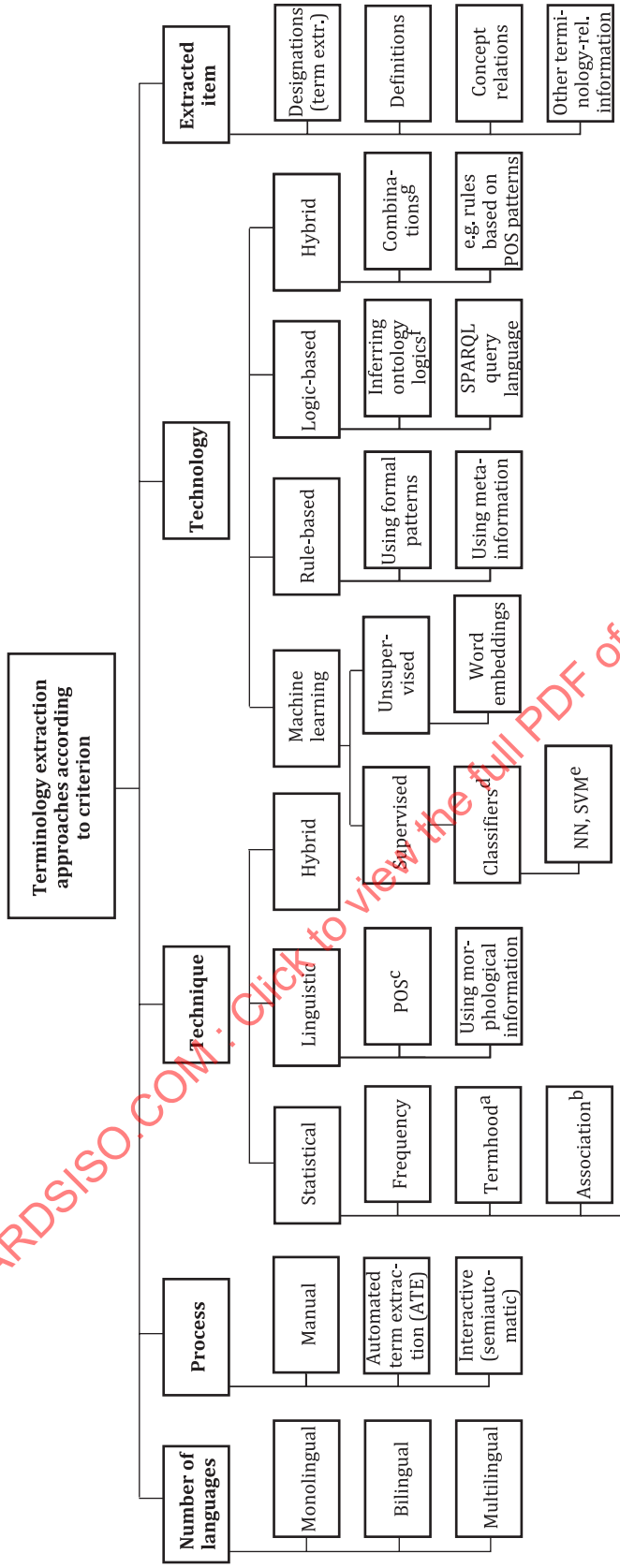
NOTE When creating a bilingual or multilingual corpus of texts and their translations, to aid the term extraction process, an alignment tool can be useful for creating translation memory exchange (TMX) files or bitexts. Sometimes corpus alignment tools align better than the term extraction tool's built-in alignment algorithm. More effective alignment improves the results of bilingual term extraction, because the extraction tool will consider the correct segments for equivalent candidate terms.

In order to make the terms specific to a subject-field text corpus stand out in the candidate term list it generates, some extractors will compare the results of the extraction from the selected text corpus with those extracted from a reference corpus.

4.4 Terminology extraction approaches and methods

4.4.1 Classification of terminology extraction approaches

Figure 1 depicts a range of terminology extraction approaches, each corresponding to a leading criterion, that are further subdivided into terminology extraction methods.



[a] Degree of termhood, e.g. log likelihood (LL).
[b] Degree of association, e.g. pointwise mutual interest (PMI) or chisquare.
[c] Filtering by grammatical categories or patterns (POS), e.g. noun phrase extraction.
[d] Classifiers (e.g. term/no term) learning from annotated data sets.
[e] Neural networks, support vector machines (SVM), decision trees.
[f] Inferring logics of ontologies (description logic, first order logic).
[g] Combinations or sequences of the other methods.

Figure 1 — Classification of terminology extraction approaches

Approaches to terminology extraction can be structured according to:

a) number of languages;

b) process;

c) technique;

d) technology;

e) extracted item.

Within the chosen approach, several terminology extraction methods can be applied, for example:

— statistical;

— linguistic;

— machine-learning-based;

— logic-based;

— rule-based.

Terminology extraction approaches and methods are detailed in 4.4.2 to 4.4.6.5.

4.4.2 Extraction method according to the number of languages

4.4.2.1 General

Extractors can focus on one language, two or more.

4.4.2.2 Monolingual

Some extractors focus on only one language. Others can process a variety of languages, but still only one at a time.

4.4.2.3 Bilingual

Bilingual terminology extraction tools can extract terms from a source text and a target text, one pair of languages at a time. These texts are compiled in a text corpus. Bitexts are ideal for mining by bilingual terminology extraction tools. Few tools can handle comparable corpora (in which original language texts on the same topic are collected in both languages). A bilingual extractor can simply present equivalent text segments for the user to consult to locate an equivalent term, or even use an algorithm to propose a possible equivalent for validation.

4.4.2.4 Multilingual

Parallel texts in more than two languages are rarely available for terminology extraction. Therefore, constructing multilingual terminology resources frequently depends on the compilation of multiple bilingual terminology extraction outputs.

4.4.3 Extraction method according to the process

4.4.3.1 Manual terminology extraction

The most basic form of manual terminology extraction involves highlighting relevant terminological data in a text and manually copying it into a list or termbase.

4.4.3.2 Automated terminology extraction

Automated terminology extraction generates a list of candidate terminological data for the user to validate later, but does not allow validation during this phase.

4.4.3.3 Semi-automated terminology extraction

With semi-automated terminology extraction, the user can validate candidate terminological data within the tool, before export. Validation at a pre-export stage reduces the noise in the resulting exported data.

4.4.4 Extraction method according to the underlying technique

4.4.4.1 Application of approaches

Most of the approaches described in this subclause apply primarily to term extraction unless otherwise stated.

4.4.4.2 Statistical terminology extraction

4.4.4.2.1 Overview

Statistical term extraction counts occurrences of candidate terms. The results are used to assess aspects, including:

— the frequency of terms in a text corpus;

— their degree of relevance to a given subject field (termhood);

— the degree of association of words in an n-gram (unithood).

Statistical term extraction relies on the relative frequency of tokens in a text corpus or their distribution within this text corpus.

Specific processing steps of statistical term extraction include:

— identification of relevant parameters such as token frequency, correlation and size of text corpus/ reference corpus;

— calculation of metrics based on selected formulae.

Depending on the type of targeted statistical data, standard tools or customized programs are used.

4.4.4.2.2 Frequency: counting occurrences

Frequency statistics are the basis for many terminology decisions and for more complex terminology extraction methods. These statistics usually count types rather than tokens and include multiword terms.

To obtain useful results using this method of term extraction, the text corpus can be filtered to exclude certain high-frequency words (the stop words containing little semantic information, such as articles, conjunctions, prepositions and auxiliary verbs). Although it can be efficient to exclude very-low-frequency words which do not reach a defined threshold for relevancy, there is a risk that some relevant terms will be overlooked. Depending on the project goal, an approach other than merely statistical can be considered.

Lexical units (single-word or multiword lexical units) that are used as terms in subject-field-specific texts commonly occur more frequently in subject-field-specific text corpora than in general-language usage. Consequently, identifying candidate terms involves comparing the frequency of lexical units in a subject-field-specific text corpus to reference corpora (a general-language text corpus or another subject-field-specific text corpus). If a lexical unit's frequency in the subject-field-specific text corpus is significantly higher than that in the reference corpus, it is very likely to be a term.

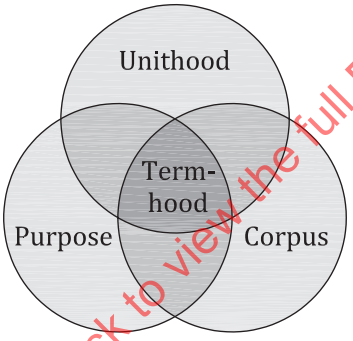
4.4.4.2.3 Termhood: relevance of terms

Termhood comprises the intersection among the following (see Figure 2):

— unithood (see 4.4.4.2.4);

— usage in the text corpus on which terminology extraction is applied;

— purpose of the terminological data collection.



NOTE Source: ISO 26162-3:2023, Figure 1.

Figure 2 — Termhood

Apart from frequency, the dispersion of candidate terms across various types of textual materials available within an organization's text corpus plays a crucial role for termhood.

The convergence of high statistical frequency and dispersion throughout a text corpus is often expressed as keyness and is viewed as an indicator that a lexical unit can indeed be considered a term in the subject field in question.

Different types of metrics are used to calculate the termhood of candidate terms, e.g. the log likelihood ratio, the Chi square ratio or the Jaccard coefficient (see Reference [17]).

4.4.4.2.4 Unithood: degree of association of words

Multiword lexical units that show relatively stable syntagmatic structures and recur with collocational frequency in an organization's text corpus feature a high degree of unithood. They are therefore used in communicative structures demanding consistency and qualify as terms when extracted for a specified purpose, see Figure 2.

A multiword lexical unit which meets the criteria for unithood functions as a term if it designates an identifiable concept in the textual and operational context in question.

The metrics used to determine the termhood of a group of words measure the significance of association between term elements. They compare the frequency of the occurrence of a phrase with the frequency of the occurrence of its components.

A wide range of term extraction procedures rely on “mutual information”, which is a concept adopted from information theory that involves measuring the statistical interdependency of two words. Since the words that form a given word pair can occur at a certain distance from each other, a suitable window is defined (e.g. four words each to the left and right of the semantic head of the compound), and mean and deviation are calculated. The mean is calculated from all the distances between the two words in the window range, while the variance measures how far the individual word distances deviate from the mean (see Reference [17], pp. 158–159).

Collocations are characterized by a very low deviation of the word distances from the mean, while a high deviation indicates the independence of the two words from each other.

4.4.4.3 Linguistic terminology extraction

Linguistic terminology extraction methods are specific to individual languages and more precise than statistical ones.

Linguistic terminology extraction relies on grammatical and morphological features of text corpus content, especially lexical units. In languages for special purposes (e.g. in Indo-European languages), terms often occur as single nouns. In Germanic, Nordic and Slavic languages, they form compound nouns or multiword units consisting of adjectives and nouns. This pattern varies in Romance languages, where prepositional phrases often replace noun-adjective strings. Character-based (e.g. Chinese) or root-based (Arabic) languages have their own patterns for combining characters and roots, with the result that the relationship of terms to “words” per se varies from language to language. In a language like Turkish, certain morphemes and affixes are used to specify the meaning of words within the context of language for special purposes. Furthermore, specific prefixes and suffixes are prevalent in some subject fields, and different languages frequently have their own patterns for adding functional and elisional elements to form compounds. The absence of white space in some languages further complicates the task of automatic term recognition. Solutions that work for one language do not always work for another, and linguists in some languages (e.g. Chinese) have created extensive resources to facilitate text corpus management and term recognition.

For details on term formation and designation patterns, see ISO 704:2022, Annexes B and C.

Linguistic terminology extraction relies on grammatical features of text corpus content, especially lexical units. In languages for special purposes, terms often occur as single nouns, compound nouns or multiword units consisting of adjectives and nouns. Furthermore, specific suffixes are prevalent in some subject fields (e.g. the suffix -itis in medical contexts).

Before commencing linguistic terminology extraction, it is necessary to perform a linguistic analysis of the text corpus content. Morphological analysis is required for identifying subject-field-specific suffixes, whereas part-of-speech (POS) tagging annotates the respective word class to each token and thus represents the prerequisite for finding syntactic patterns underlying term formation.

Processing steps include:

— morphosyntactic analysis (including POS tagging) of the text corpus;

— defining patterns for relevant terms;

— filtering candidate terms based on these patterns.

Since a linguistics-based approach can involve many aspects of the language, either a set of dedicated tools or custom programs are used.

The programs used should have features applicable to the language(s) of the terminology extraction project. For example, POS taggers use models learned from annotated data in a specific language, i.e. a POS tagger will use a different model for English and for French.

POS taggers use pre-defined probabilistic heuristics for determining which part of speech is most plausible for a certain token. To optimize results, POS taggers are usually trained using a small, manually annotated sample (approximately 10 %) of the text corpus to be analysed before they are applied to annotate the larger part of this text corpus.

Definition extraction has evolved much more slowly than term extraction and is highly language-dependent. One method of extracting definitions for concepts designated by terms is identifying relations between two terms within a sentence. For example, the so-called Hearst pattern[16] "A is a B" indicates a hierarchical relation between the nouns A and B, where A represents a subordinate concept of B and B represents the superordinate concept of A.

Extracted terms can also serve as a basis for bootstrapping approaches, that is, as a starting set for identifying further terminology and to retrieve more texts to include in the corpus as well. An example for resources on the web which feature synonyms to the identified term are lexical databases that link, by semantic relations, the concepts or contexts underlying a word (see WordNet[18] and FrameNet[14]). Due to explicitly modelled hierarchical relations, ontologies can be useful resources in order to identify terms for concepts subordinate to a concept labelled with an already extracted term.

4.4.4.4 Hybrid terminology extraction

Hybrid terminology extraction often combines useful procedures from both statistical and linguistic terminology extraction.

4.4.5 Extraction method according to the underlying technology

4.4.5.1 Machine-learning-based terminology extraction

4.4.5.1.1 General

Machine-learning-based terminology extraction is a process of automatically identifying and extracting terms from textual data using machine-learning algorithms implemented on neural networks.

When setting up a neural-network-based terminology extraction, specific attention should be paid to the following aspects:

— identification of a relevant text corpus that reflects the characteristics of the performance of the intended model in terms of languages, subject field and language register;

— specification of a standard annotation scheme and annotation guide comprising the relevant decisions regarding the types of terms, their scope and possible embeddings;

— provision of enough data in relation to the number of parameters of the neural model in order to guarantee the actual convergence of the model and avoid overfitting;

— definition of quality criteria for the annotated data that ensures its homogeneity;

— clear identification of two separate sets of data for training and testing the model.

The quality of the results of machine learning strongly depends on the quality of the data used for training. Therefore, the first step is to ensure that the data are not biased, are evenly distributed, and are representative of the subject field.

Often, there is more than one method for training a model, and it is good practice to compare the results of relevant algorithms to select the most promising approach.

In most cases, machine-learning methods require some degree of custom programming.

Standard practice is to divide the available training material in two or three uneven parts: a training set, a validation set and a test set. The model is created based on the training set. The validation set supports the training by allowing different parameter settings to be tested, and the test set comprises data that the model has not seen and which can be used to measure the performance of the model.

Terminology extraction based on machine learning can be divided between supervised and unsupervised machine learning.

Some of the output of machine-learning-based terminology extraction include:

— word embeddings (unsupervised), which help obtain semantic information about terms and recognize similarities between terms (monolingual or cross-lingual);

— terminology recognition model (supervised).

4.4.5.1.2 Supervised machine learning

Supervised machine learning has two phases. First, a model is "learned" based on annotated data. The resulting learned model is then applied to new, unknown data and outputs a result (e.g. a list of candidate terms). This learning is accomplished using a neural network that calculates the probability of a result (e.g. a word being classified as a candidate term or not) based on the features available in the input data. A feature can be a token or a word order. The complexity of the network and its layers depends on the algorithm chosen and the expected output.

Supervised learning approaches can be fine-tuned with a number of parameters. The text corpus consists of two basic parts: the data as such (e.g. a collection of sentences) and the desired output (e.g. a category).

Neural network methods constitute supervised machine-learning methods based upon computational models organized as a graph of elementary combinatory units (neurons) that are organized in layers and combine their inputs to compute an output. This output is in turn transmitted as an input to the neuron units of the next layer. Training data created from input and output reference pairs that reflect the intended system behaviour are used to set the neural network model parameters.

4.4.5.1.3 Unsupervised machine learning

Unsupervised machine learning analyses raw text corpus data and tries to recognize the relevant features needed to describe the data, to cluster it or to predict an output. Examples of output include word embeddings, which can be used to extract semantic information.

Vector-based term recognition metrics stem from the information retrieval (IR) field, which relies on vector space models. In IR, a search query is understood as a vector, that is a directed quantity, which results from the words used in it and their number. The possible matched searches in the form of result documents are thought of as individual points in a high-dimensional, metric vector space. The more similar the result vector is to the query vector, the more accurate the results will be.

Here, similarity is understood as the presence of as many identical features as possible – related to term extraction: words – in both vectors. In addition, the individual words receive a relevance weighting, depending on their frequency of occurrence in a document and in the entire text corpus (the so-called "TF-IDF measure"), in order to eliminate stop words and thus superfluous dimensions in the vector space. The similarity is very high for parallel query and result vectors, especially when the cosine distance between the two vectors is small (see Reference [19]).

This approach is very productive not only in classical information retrieval: it can be transferred from the underlying document-word matrix used in document indexing to a word-word matrix and to a modifier-head matrix that are well suited for identifying conceptual similarity of multiword units (see Reference [17], p. 297).

The great advantage of vector-based methods, which also form the basis for clustering words into specific word fields, is that they take into account paradigmatic relations between terms such as synonymy, hypo- and hyperonymy as well as meronymy (compare Reference [20], p. 30).

4.4.5.2 Rule-based terminology extraction

Terms can be extracted from a text corpus according to specific rules. For example, the terms can match specific regular expressions or specific tags in an extensible markup language (XML) file.

The first step in this process is to identify the patterns that are most likely to produce the anticipated terminology results. These patterns can be purely formal (e.g. by identifying acronyms as a group of up to *n* capital letters). Such patterns can be based on metadata embedded in the file content (e.g. tags used to identify index entries). They can also be combined with linguistic information (e.g. part of speech or morphemes).

4.4.5.3 Logic-based terminology extraction

The data used to extract terminological information by logic-based terminology extraction shall reveal a structure suitable for logical queries. These data are characterized by objects, attributes and relations between objects. Logical queries use this structure to infer new information that can be used in the identification or creation of candidate terms, as well as in the identification of concept relations.

Data sources can include internal data repositories of organizations or can be publicly available data such as DBpedia[13] or other available Linked Open Data (LOD) sources that are available mainly, but not only, in Resource Description Framework (RDF) or Web Ontology Language (OWL).

Since relations and attributes vary depending on the data source, the relations and attributes to be used in logical queries shall first be identified. Such attributes can include, for example, attributes that support the classification of concepts, that help to identify instances of a class, or that refer to definitions and synonyms.

Query languages or custom programming can be used to integrate logic-based extraction technology into a terminology extraction pipeline.

4.4.5.4 Hybrid terminology extraction

Hybrid models rely on complex architectures that combine several of the above-mentioned technologies. Such systems ideally integrate the relevant constraints for each approach and ensure the coherence of the resulting design.

4.4.6 Extraction method according to the extracted items

4.4.6.1 General

Among terminological data that can be extracted during the process of terminology extraction, it is also desirable to extract not only terms, but also definitions (to clarify the concept the term represents), concept relations (to provide an overview of the concept system based on generic, partitive or associative relations) as well as other useful information for a terminology record. The richer the terminological data included in the concept record, the more useful the terminological resource.

4.4.6.2 Criteria for extracting terms

There are several criteria that can be used to extract terms from a text corpus, depending on the subject field and purpose of the terminology extraction project. In addition to termhood and unithood (see 4.4.4.2.3 and 4.4.4.2.4), the criteria in this subclause can be considered.

Critical to any type of terminology extraction task is the proper delimitation of terms. Frequent proximity between words is insufficient to create termhood. In a list of candidate terms, several strings of words can appear frequently together in a text corpus. The strings “Multigenerational Home Renovation Tax Credit”, “remaining Employment Insurance costs relate” and “closely tracked evolving personal opinions” can serve as an example for this observation. Of these five-word candidate terms provided by an extraction tool, only the first is a term and already properly delimited. The second requires term delimitation work by a human to glean the term “Employment Insurance” from the longer candidate term. The last string contains no term.

Term delimitation is also particularly important with regard to multiword terms. Within some longer multiword terms, shorter terms can be found. For example, within the term “automatic gear-box selector dial” designating one concept, the terms “automatic gear-box” and “selector dial” for two other concepts can be found.

4.4.6.3 Criteria for extracting definitions

To find and extract definitions, it helps to look for definition-specific patterns. For example, searching an English text corpus for the strings “is defined as”, “is a” or “such as”, or a French text corpus for “défini”, “définis”, “définie” or “définies” is a first step to locate definitions. It is possible to search a text corpus manually for these patterns in some terminology extraction tools that offer text corpus search.

4.4.6.4 Criteria for identifying concept relations

The identification of concept relations can involve the recognition of syntactical patterns that describe relations among concepts. For example, a first step to identify concepts and their interrelations involves searching an English text corpus for lexical-semantic relations indicated by, for example, “is part of”, “is a kind of”, “comes before” or “is produced by”. It is possible to search running text manually for these patterns.

Furthermore, terms can be used as a start set for retrieving identical node labels in domain-specific ontologies. By retrieving the labels from those ontology nodes and from their neighbour nodes, as well as the respective relations that interlink them, semantically related concepts can be detected. Since ontologies feature explicitly defined relations between nodes, this information can help expand a concept model featuring subject-field-specific terminology. The same approach can be applied to Simple Knowledge Organization System (SKOS) taxonomies, where hierarchical relations between concepts are explicitly labelled as “broader” or “narrower”.

4.4.6.5 Criteria for extracting other terminology-related information

In addition to definitions, adding a term used in context (e.g. in a sentence) to a terminology record can provide additional information about the concept, as well as how it is used.

Also, if the term is only or most often used with particular verbs, including collocations can also be very helpful. For example, the verb to use in English with the noun “crime” is “commit”. Having this information in the entry for “crime” can help people whose first language is not English to use the term with more confidence. Some extractors offer a concordance tool such as a Key Word In Context (KWIC) tool to help locate all strings of a given term and display the context segments with the search term in the middle. Sorting by the 1, 2 or 3 words to the left or right of the term makes it easier to identify collocations and even other complex terms.

Some terminology extraction tools will automatically add a context sentence to the terminology record, while others present a set of context sentences from which the language professional can select the one(s) to be added to the record.

4.5 Term extraction output

4.5.1 Filtering candidate term lists

All candidate term lists contain a certain amount of noise (undesired candidate terms) or silence (overlooked candidate terms). Noise and silence are closely related to the concepts of recall and precision. Human intervention is always needed at some stage to obtain validated terms. This can take place either before or after the creation of draft terminology records based on the extraction results or some other application defined by the project in question. Some tools can filter by part of speech, sort by frequency, alphabetical order or number of lexical units in a candidate term. Some can go farther, so that when validating a candidate term, one can consult the text segments in which it is found, even to the point of sorting by 1, 2 or 3 tokens to the right or left of the candidate term, making it easier to find collocations.

Terminology product length limits can also affect term selection during extraction or validation of automated term extraction. If, for example, the goal is to create a 300-term glossary, then it can be sensible to discard some legitimate terms from a 1 000-candidate term list so as not to go over the limit. However, if the goal is to cover a full range of terms occurring in a text corpus, then all terms from a systematic extraction can be accepted.

When examining the candidate term list for a technical subject field, terms that are not specific to that subject field should not be taken into consideration. The appropriateness of these terms can be evaluated by comparing the subject-field-specific text corpus to a general-language reference corpus.

Project deadlines can also be a factor limiting term selection. If a deadline is short allowing little research time and a list of candidate terms is long, then prioritizing candidate terms that require little further research can be an option, either by restricting the scope of the project to key concepts or by consulting subject-field experts.

4.5.2 Assessing term eligibility

The criteria for selecting and extracting terms will also vary according to the goals and purposes of the term extraction project. In order to ensure that results align with specified goals, it is essential to determine which terms are to be extracted for a given terminology product. For example, specifications for a given use case require the creation of a termbase listing official designations for species of fish in a given jurisdiction. One approach is to perform term extraction using a corpus of maritime legal texts on catch limits. The process will extract any fish names that do occur in the text corpus. During the extraction process, the decision can be made to discard candidate terms that are legal terms or the proper names of islands. Even though these items are themselves terms, they do not fall inside the parameters for the project.

Possible term eligibility criteria (non-exhaustive) should consider the following:

— Project specification: which terms should be used or translated consistently.

— Subject-field specificity: whether to limit selection to subject-field-specific terms or to allow non-special-language words, depending on specifications.

— Term frequency: High frequency is one indicator that a term can be important or that it can be a term specific to the subject field covered by the text corpus. Focusing on terms that occur many times in a text corpus often provides good return on investment (ROI) for standardization (e.g. in translation scenarios).

— Term distinctiveness: Term incidence in subject-field-specific text corpora can be compared to their frequency in a general language text corpus. Terms that appear more frequently in the subject-specific text corpus than in the reference corpus are often of value as candidate terms, because they are more likely to be subject-field-specific.

— Term specificity: Appellations and proper names (e.g. names of people, locations and organizations) can be considered for inclusion (or exclusion) as candidate terms for extraction. They always designate either a specific entity or a type of identical objects (such as product names), so they can be relevant (or irrelevant), depending on specifications for the terminology management use case.

— Legal issues: Private companies seeking to protect a brand can extract proprietary product or process names.

— Cost savings: Unfamiliar, ambiguous or disputed terms can require significant time or resources to document over the course of a project or after publication. Extracting and documenting these terms as part of the production process avoids repeating costly research.

— Acronyms or other abbreviated forms: The best practice with abbreviated forms is to list them together with their full forms in order to prevent confusion.

— Existence of synonyms: Terms and their synonyms are especially important to record (and perhaps to designate one as preferred later) to avoid ambiguity or confusion.

— Controlled language: If controlled language is used by an organization, any term to be controlled merits extraction.

— Novelty: In emerging fields, extracting new terms (neologisms) and tracking the evolution of their usage can help keep texts uniform until usage settles.

It is important not just to consider what is desirable to extract, but also what is not. If a terminology record for a candidate term already exists, then it can be superfluous to have it appear in a list of candidate terms.