Controlled Multilingual Thesauri for Kazakh Industry-Specific Terms

Bayekeyeva, Aisaule; Tazhibayeva, Saule; Beisenova, Zhainagul; Shaheen, Aigul; Bayekeyeva, Ainur

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Bayekeyeva, A., Tazhibayeva, S., Beisenova, Z., Shaheen, A., & Bayekeyeva, A. (2021). Controlled Multilingual Thesauri for Kazakh Industry-Specific Terms. *Social Inclusion*, *9*(1), 35-44. https://doi.org/10.17645/si.v9i1.3527

Social Inclusion (ISSN: 2183–2803)
2021, Volume 9, Issue 1, Pages 35–44
DOI: 10.17645/si.v9i1.3527

Article

Controlled Multilingual Thesauri for Kazakh Industry-Specific Terms

Ainur Bayekeyeva [1,*], Saule Tazhibayeva [2], Zhainagul Beisenova [2], Aigul Shaheen [2] and Aisaule Bayekeyeva [3]

[1] Department of Translation Theory and Practice, L. N. Gumilyov Eurasian National University, 010000 Nur-Sultan, Kazakhstan; E-Mail: a_baekeeva@mail.ru
[2] Faculty of Philology, L. N. Gumilyov Eurasian National University, 010000 Nur-Sultan, Kazakhstan; E-Mails: sauletazhibaeva@mail.ru (S.T.), zhaina_b@mail.ru (Z.B.), akingan@mail.ru (A.S.)
[3] State Material Reserve Committee of the Ministry of National Economy of the Republic of Kazakhstan, 010000 Nur-Sultan, Kazakhstan; E-Mail: aisaule68@mail.ru

* Corresponding author

Submitted: 31 July 2020 | Accepted: 15 October 2020 | Published: 14 January 2021

Abstract
This article discusses the practical issues of compiling controlled multilingual thesauri for the purposes of industry-specific translation (IST). In the multilingual, transnational and globally connected Kazakhstan, IST is a much-needed translation service. IST is an interdisciplinary field between terminology, computational linguistics, translation theory and practice. Most of the professional guides, dictionaries and glossaries are systemized in alphabetical order and contain multiple variants for the terms searched. Therefore, there is an urgent need to create a systemized controlled multilingual thesaurus of industry-specific Kazakh, English and Russian terms in order to provide multilingual users with an interoperable and relevant term base. Controlled multilingual thesauri for industry-specific terms are the most effective tools for describing individual subject areas. They are designed to promote communication and interaction among professionals, translators and all Automated Information System users of specific fields irrespective of their location and health conditions. Unlike traditional dictionaries, controlled thesauri allow users to identify the meaning with the help of definitions and translations, relations of terms with other concepts, and broader and narrower terms. The purpose of this research is to unify and systematize industry-specific terms in Kazakh, to provide Russian and English equivalents, and to classify the terms into essential rubrics and subjects. Based on the Zthes data scheme to create a controlled multilingual thesaurus of industry-specific terms, the major rubrics have been formulated, and about 10,000 Kazakh mining and metal terms approved by the Terminological Committee of Kazakhstan have been structured.

Keywords
controlled thesaurus; controlled vocabulary; industry-specific terms; interoperable thesaurus; Kazakhstan; multilingualism; multilingual thesaurus; terminology; thesaurus

Issue
This article is part of the issue “Social Inclusion and Multilingualism: The Impact of Linguistic Justice, Economy of Language and Language Policy” edited by László Marácz (University of Amsterdam, The Netherlands / L. N. Gumilyov Eurasian National University, Kazakhstan) and Zsombor Csata (Babeș-Bolyai University, Romania / Hungarian Academy of Sciences, Hungary).

1. Introduction

1.1. Language Situation in Kazakhstan

Kazakhstan is a multiethnic, multilingual and multicultural country. During the twentieth century, 126 ethnic groups found residence in Kazakhstan. These ethnicities migrated to Kazakhstan from different regions of the Soviet Union for various reasons: (1) deportation during the collectivization period in the USSR in the 1920s; (2) deportation during the period of the Second World War; (3) industrialization during the Second World War;

and (4) the Virgin Land Campaign in the 1950s and 1960s (Nevskaya & Tazhibayeva, 2015a).

Many people from various populations came to Kazakhstan's cities to work in the factories, in state and municipal administrations, in organs of the Communist Party and other fields. Such migration volume strongly influenced the ethnic-demographic, social and language situation in Kazakhstan. From a monolingual country in the early 1920s, it turned into a bilingual one, with a strong Russian language dominance.

As a result, the Kazakh language was the language of rural areas, where the majority of Kazakh population lived. The language situation in Kazakhstan changed after the collapse of the USSR. The subsequent years of Kazakhstani independence have had a positive impact on the state language by broadening the usage of Kazakh in education, mass media, and culture. The promotion and full functioning of Kazakh as well as the policy of multilingual education was announced in the State Program on Education Development for 2011–2020. A new generation of young Kazakhs carries out their activities in Kazakh, Russian and English. However, code shifting into Russian is very frequent even now and Russian serves as an intermediate language when providing Kazakh–English translation.

1.2. New Challenges to Kazakh Industry-Specific Terminology

Scientific and technological progress in the 1960–1980s, globalization since the 1990s, and the world pandemic situation in 2019–2020 have created new challenges and opportunities in information exchange for the emerging new field of industry-specific translation (IST). Globalization in a multilingual world, the 4th Industrial Revolution and the recent world pandemic have impacted Kazakhstan's economy, industry and other related fields such as translation services. Seventy percent of Kazakhstan's state bodies and ministries, 80 percent of educational establishments, schools and universities, and 100 percent of disabled employees and other vulnerable communities have been changed over to home offices (Republic of Kazakhstan, 2020).

The prevalence of linguistically diverse and multilingual societies are increasing due to globalization and Europeanization (Marácz & Adamo, 2017; Nevskaya & Tazhibayeva, 2015b). English often is considered to be the lingua franca in international communication, business and technology. Therefore, translation from/to English still remains an imperative in most non-English-speaking countries of the world, and trilingual Kazakhstan is no exception (Yeskeldiyeva & Tazhibayeva, 2015). Multilingual IST is a cross subject and multidimensional field involving terminology, computational linguistics, translation theory and practice.

As a branch of terminology, bilingual industry-specific terms in Kazakh and Russian were first unified in the 1950s by the initiative of Kazakh scholars Kanysh Satpayev and Alimkhan Yermekov (Kydyralina, 2014). The role of the Kazakh language in texts of science and technology remains in the shade leaving Russian to be the dominant language. In IST Russian serves as an intermediate language. The lack of English–Kazakh and Kazakh-English dictionaries in subject areas is a serious problem for the professionals and translators. Furthermore, when searching for the necessary terms in online dictionaries, the absence of trilingual online dictionaries creates difficulties especially for vulnerable communities. Rapidly growing economic cooperation in Kazakhstan make it necessary to enrich the term base with various new terms denoting technology and equipment. Direct English–Kazakh translation occurs very rarely. Creating a new generation of dictionaries will enhance opportunities for Kazakhstan's translation services and strengthen the Kazakh language status as a state language.

The aim of our controlled multilingual thesaurus is to offer a way for all AIS users to find the relevant data related to the industry-specific terms without having to search several dictionaries for the relevant information and accurate terminology, and thus avoiding misunderstandings and mistranslations.

1.3. Kazakh Industry-Specific Dictionaries

Automated Information System (AIS) users and industry-specific translators in Kazakhstan's ministries and other state bodies use a variety of online dictionaries and language tools, particularly *ABBYY Lingvo*, *FreeDict*, the *Free On-line Dictionary of Computing*, *WordNet*, among others. The most widely used online dictionaries and language tools in the multilingual Kazakhstan are *Sozdik*, *Termincom* and *Multitran*, which meet the requirements of bilingual dictionaries. In order to provide translations in three languages, the AIS users and translators have to consult two bilingual dictionaries as *Sozdik* and *Termincom* for Kazakh and Russian, then *Multitran* and *ABBYY Lingvo* for Russian and English. When searching for the necessary terms in online dictionaries, the absence of trilingual online dictionaries creates difficulties especially for vulnerable communities.

Many trilingual industry-specific dictionaries are available only in hard copies and in the current pandemic situation, our controlled multilingual thesaurus gives equal opportunity to all AIS users to access these dictionaries irrespective of their location and health conditions. Another issue in terminological dictionaries is multiple variants of a certain term. Industry-specific terms have widely a used synonymic range of terms and the problem of choosing the relevant term between the ranges of synonyms often arises. For example, terms such as 'beneficiation,' 'concentration,' 'dressing,' 'enrichment,' 'refinement,' 'separation,' 'treatment' and 'washing' for the Kazakh *кен байыту* and Russian *обогащение*.

In this article, the selected mining and metal terms, approved by the Terminological Committee

of Kazakhstan have been analyzed by an expert of Kazakhstan's Ministry of National Economy with special attention for use by the disabled, translators and terminologists. Moreover, a unified search tool on Zthes data scheme has been demonstrated to provide relevant information with the examples of industry-specific terms in trilingual context. The article has been written by a group of authors with the support of the State Material Reserve Committee of the Ministry of National Economy of the Republic of Kazakhstan to support developing industry-specific terminology with the involvement of experts from vulnerable communities.

2. Controlled Multilingual Thesaurus in Digital Library

2.1. Thesaurus: General Information

The thesaurus or ideographic dictionary in general (in Greek *Θησαυρός*, in English 'treasure') is a collection of information, a corpus of concepts, definitions and terms of special fields of knowledge or industries, with examples of their use in a context. Today, one of the most well-known thesauri was compiled by the British lexicographer Peter Marc Roger and published in 1852. The original name of this thesaurus was the *Thesaurus of English Words and Phrases*.

In the field of machine translation, Masterman (1961) was the first to utilize a thesaurus in 1961. Masterman's thesaurus contained 15,000 concepts defined as the basic vocabulary. With the help of thesauri, correspondence was established between the language of user requests and documents in the information system. Shreider (1965) proposed to consider the thesaurus as a system of knowledge reflected in the language, in which case the thesaurus itself becomes interesting and not just an auxiliary tool. Other controlled thesauri are the *WORDNET* thesaurus for English, developed at Princeton University, its analogue *RussNet* thesaurus for Russian, developed by the Mathematic Linguistics Department at St Petersburg State University, the multilingual *UNESCO Thesaurus*, *NISO* and *LOGOS* (see ANSI/NISO, 2003, 2005; IFLA, 2009).

According to the definition of the International Organization for Standardization (ISO), a thesaurus is a dictionary of managed indexing language, formally organized in order to establish explicit a priori relationships between concepts (ISO, 1985, 1986). This definition establishes lexical units and semantic relationships between these units as elements that constitute the thesaurus. Thesaurus relations such as genus-species, part-whole, etc., are imposed on the taxonomy structure, i.e., they identify the main taxonomy of the subject area.

Legislative Indexing Vocabulary designers of the Library of Congress Linked Data Service (id.loc.gov) recommend the following rules: (1) Terms of thesaurus should represent the concepts that are actually mentioned in the source documents and should be selected from considerations of their use efficiencies in searching for documents; (2) an important factor in including the term is the frequency of its mention in the texts that must be checked periodically; (3) including new terms in the thesaurus should take into account already included thesaurus terms (Library of Congress, n.d.).

Thus, the candidate terms should be checked for consistency with respect to their generality and specific use with other terms of the thesaurus. It also should be checked as to whether or not the candidate term represents a separate concept that does not have correspondences among the existing terms in the thesaurus. It is necessary to avoid including terms whose values overlap with the values of already existing terms in the thesaurus because it might be difficult for indexers, users and translators to differentiate between these terms.

2.2. Controlled Multilingual Thesaurus of Industry-Specific Terms in Kazakh

The goal of developing the controlled multilingual thesaurus is to unify and systematize industry-specific Kazakh terms in conformity with the requirements of modern interoperable thesauri to be easily integrated with other thesauri, to describe and provide Russian and English equivalents, and to classify the terms into relevant topics using related rubrics and associated relations in a hierarchical order. The controlled thesauri provide a special type of vocabulary with general and specialized terminology, in which relations such as broader and narrower terms, synonyms, antonyms, paronyms, hyponyms, hypernyms etc. are indicated with their equivalents in other languages.

Based on the study of standards and various approaches to controlled thesauri, the Zthes data scheme was chosen to create a controlled multilingual thesaurus for industry-specific terms. Within the framework of this research, major rubrics have been created for about 10,000 Kazakh terms on mining and metal industry. These are 'mining and metals,' 'minerals,' 'exploration,' 'production,' 'equipment' and 'technology.' The use of multilingual thesauri in digital libraries is effective because of their interoperable characteristics, so this multilingual thesaurus can be easily integrated into international databases such as the UNESCO, NISO and LOGOS thesauri.

The proposed thesaurus model was implemented by IRIS, an integrated resource information system developed by the Institute of Computational Technology of the Siberian Branch of the Russian Academy of Sciences (SB RAS; Shokin, Fedotov, Zhizhimov, & Fedotova, 2015; Tussupov, Sambetbayeva, Fedotov, Fedotova, et al., 2016). The Zthes data scheme was chosen to compile the multilingual thesaurus for industry-specific Kazakh, Russian and English terms.

To date, the most effective way to solve the lexicographic problems in IST was to organize the relevant information into information systems provided by digital libraries. Digital libraries offer specialized technol-

ogy to work with digital information, forming a new class of information systems designed to manage information resources (Tussupov, Sambetbayeva, Fedotov, Sagnayeva, et al., 2016). The standard approach to systemize the data is by classifying documents using taxonomies. Taxonomy is a subject classification that groups terms into a controlled thesaurus and organizes the dictionaries into hierarchical structures. To describe a subject area, usually a certain set of key terms is used, each of them denotes or describes a concept from a given subject area (Salton, 1979).

During the classification process, the relevant concepts and key terms are determined, and 'parent-child' type relations are established. The problem in providing scientific and educational information systems for industry-specific terms is that the technologies for classifying and systematizing information developed by libraries and archives do not effectively work now because of the thematic proximity of the classified documents. The efficiency in information retrieval systems to support scientific and educational activities directly depends on the use of specific thesauri.

2.3. Structure of Multilingual Controlled Thesaurus for Industry-Specific Terms

Our multilingual controlled thesaurus for industry-specific terms has been developed as a joint project between L. N. Gumilyov Eurasian National University, Kazakhstan, and the Institute of Computational Technology of the SB RAS (see Figure 1).

The aim of our controlled multilingual thesaurus is to offer a way for all AIS users to find the relevant data related to the industry-specific terms without having to search several dictionaries for the relevant information and accurate terminology, and thus avoiding misunderstandings and mistranslations. For example, the Kazakh term *ұңғыма* and Russian *скважина* has several linguistic equivalents in English such as 'bore,' 'borehole,' 'hole' and 'well.'

According to the dictionary of the Well Drillers Association (n.d.): "Typically a borehole is drilled by machine and is relatively small in diameter. A well is usually sunk by hand and is relatively large in diameter." The term 'well' is used only in the oil and gas industry because of the liquid nature of oil, while 'borehole' is used for mining and the metallurgical industry because of the solid nature of the ore. However, in Kazakh and Russian, this word is used with respect to only one meaning since *ұңғыма* and *скважина* that have no specific synonyms.

This controlled multilingual thesaurus consists of a variety of dictionaries, including a reference dictionary, multilingual dictionary, a dictionary of synonyms, a dictionary of antonyms, a dictionary of homonyms, a glossary, etc. The terms are clearly defined semantically and functionally, their linguistic equivalents are established and they are classified hierarchically.

This multilingual, controlled and interoperable thesaurus is composed of the following elements (see Figure 2):

1. A list of terms.
2. The relations between the terms, indicated by their hierarchical relative position. BT (broader term) is a connection with the parent term, e.g.: 'alloy'—term; 'metal'—broader term.
3. NT (narrower term) is a connection with a child term. BT ↔ NT communication is mutual and reciprocal, e.g.: 'brass'—narrower term.
4. UF (used to mean) is reciprocal feedback of USE, USE ↔ UF.
5. RT (related term) is a link defining a related term.
6. LE (linguistic equivalent) is a connection between linguistically equivalent terms in Kazakh, English and Russian, e.g.: *сплав* and *қорытпа*—terms in Kazakh and Russian.
7. TT (term type) is a top-level term, i.e., a term that has no related terms of a wider class (terms with the type of BT connections).

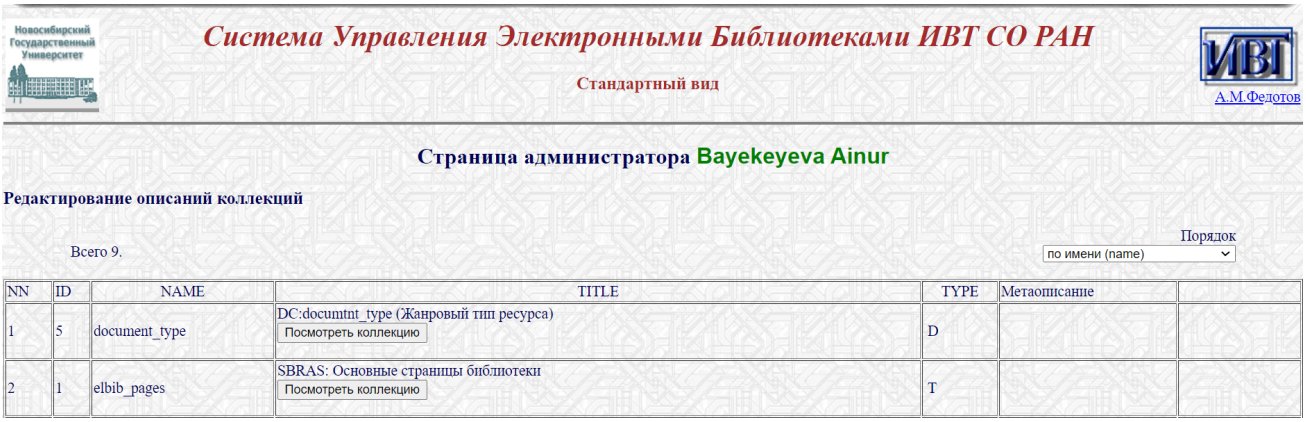

Figure 1. Homepage of the website *Control System for Digital Libraries* of the Institute of Computational Technology of the SB RAS.

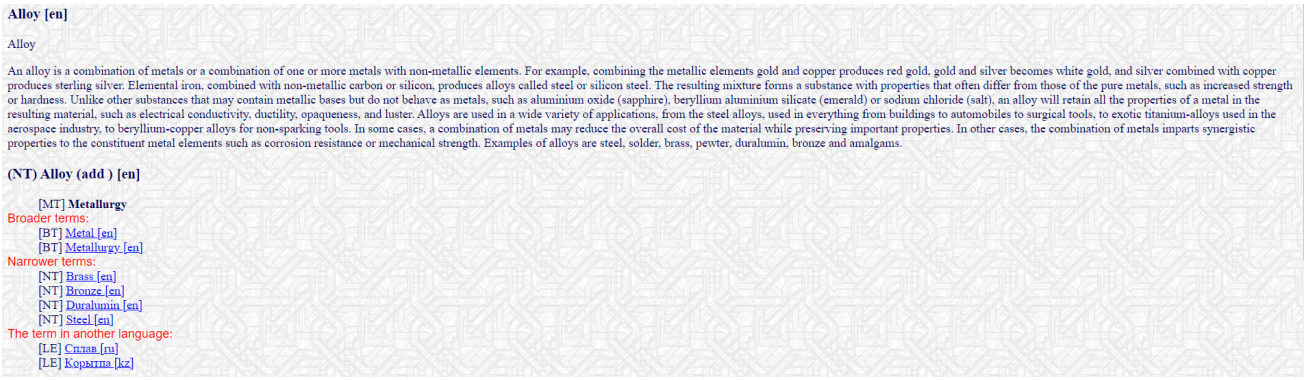

Figure 2. Hierarchical relative position of our controlled multilingual thesaurus.

8. NT (relation type) is not a top-level term, a descriptor that has BT type connections.

Most controlled interoperable thesauri are usually monolingual. The multilingual thesaurus based on a digital library is a system for managing structured catalogued collections of dissimilar digital resources, i.e., industry-specific terms and their associated relations, links and parallel texts. It provides not only a comprehensive search and navigation function through catalogues, but also shows all the users the relevant resources such as publications, documents, pictures, fact descriptions, etc.

Our multilingual thesaurus gives comprehensive and relevant information related to the terms. Some examples are: 'concentration,' used for concentrating the chemicals in laboratory; 'refinement,' used only for gold and silver refining in mining and metallurgy (but oil can also be refined); 'separation,' the exact process of separating the different metals in the ores; 'washing,' the initial process of separating the ores; 'dressing,' used for treating the ore in the special equipment; 'treatment,' which is the general name for processing the ores of any type. Only 'enrichment' and 'beneficiation' can be used synonymously for the Kazakh *кен байыту* and Russian *обогащение*.

The multilingual thesaurus based on Zthes data scheme allows all AIS users to search for the relevant information related to the specific term in a multilingual way. In the case of searching for information related to the Kazakh term *жез*, one may use *Sozdik* for the Kazakh–Russian translation, and *Multitran* or other dictionaries to translate the term into English, in addition to other dictionaries to get more information about the specific term.

2.4. Advantages of Multilingual Thesaurus

The main advantage of the multilingual thesaurus in the Digital Library is the targeted or relevant search of terms in semantic and functional aspects. The industry-specific terms in a particular field are clearly defined and linguistic equivalents in Kazakh, Russian and English are established and hierarchically classified. Thus, one can find the relevant information about the Kazakh term *жез*: Its linguistic equivalent in Russian is *латунь* and in English 'brass.' This term (*жез*) directly associates with the broader term *қорытпа* and its English equivalent 'alloy'—and subsequently with other types of 'alloys,' 'metals' and 'ores.' Every term has relevant descriptions with associative relations that allow the user to search for more information related to the industry-specific terms (see Figure 3).

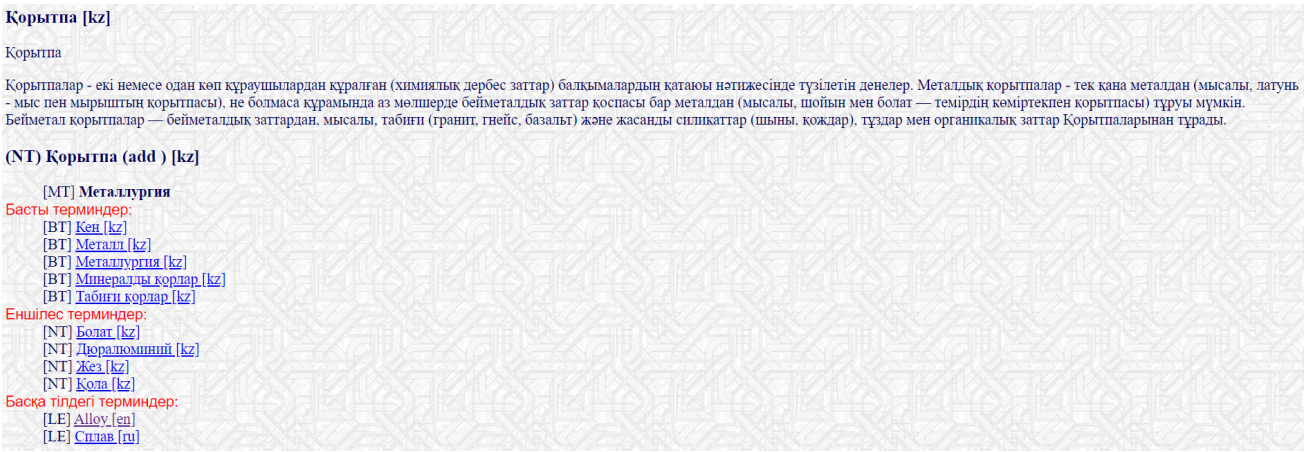

Figure 3. Example of searching for industry-specific terms.

Consequently, the controlled multilingual thesaurus in the Digital Library is an effective tool not only for professionals of specific industries, but it is useful for all AIS users regardless of their location and health conditions. This also lays the groundwork for the future development of a Kazakh language corpus, and for unifying and systemizing the industry-specific Kazakh terms in a multilingual way.

3. Controlled Multilingual Thesaurus for Information Retrieval

The information retrieval thesaurus is a normative and controlled dictionary of key terms in natural language with explicitly indicated semantic relations between terms. It was designed to describe the content of documents and search queries (ISO, 2011, 2013). The thesaurus is intended to describe a specific subject area, each term of which denotes or describes a concept from a given subject area.

This thesaurus is constructed to describe the vocabulary of descriptive Information Retrieval language, the lexical units of which are descriptors. The terms, descriptors, keywords and associative relations are the notions used in compiling multilingual thesaurus aimed at information retrieval. The selected terms in a controlled multilingual thesaurus serve as descriptors that denote certain concepts of a specific subject area and satisfy the principles of common usage, prevalence, brevity and terminological accuracy proposed by domestic and foreign terminologists (Aitbaiuly, 2013; Beisenova, 2011, 2014; Cabrė, 1998; Kaidarov, 1993; Kurmanbaiuly, 2014).

Keywords, as a separate terms, words or phrases, extracted from the text of the subject area data and document reflect the main content when indexing. The group of conditionally equivalent keywords unites not only the terms that are recognized as synonyms in the multilingual term base, but also those that can be considered conditionally equivalent in terms of information retrieval within the framework of the multilingual scope, i.e., in Kazakh, Russian and English. The associative relations express the class-type, whole-part relations between the terms of information retrieval language. These terms are stable for each subject area and indexed in the multilingual thesaurus. This can be illustrated by the industry-specific terms such as 'trammeling,' 'crushing,' 'grinding,' 'flotation,' 'filtration' and 'magnetic separation,' which are indexed in the rubric Enrichment of the subject field Mining and Metals (see Table 1).

In the information retrieval thesaurus, the following relations are indexed and applied: (1) class-type relation; (2) equivalence or synonymic relation; (3) associative relation. Source: Bayekeyeva (2018) and Bayekeyeva, Tazhibayeva, Shaheen, Beisenova, and Mamayeva (2020).

The class-type relation links two descriptors if the scope of the concept corresponding to one of the descriptors includes the scope of the concept of another descriptor, e.g.: crushing department—crushing workshop*;* crushing unit—crushing machine; crushing type—medium crushing, fine crushing, etc. (see Table 2).

The equivalence or synonymic relation occurs when searching for one of the conditional or true synonyms that allows finding in the AIS database the documents to

Table 1. The examples of multilingual descriptors.

Descriptor in English	Descriptor in Kazakh	Descriptor in Russian
Trammeling	Тас уату	Грохочение
Crushing	Ұсату	Дробление
Grinding	Ұнтау	Измельчение
Flotation	Флотация	Флотация
Filtration	Фильтрация	Фильтрация
Magnetic separation	Магнитті сепарациялау	Магнитная сепарация

Notes: The equivalents of Kazakh and Russian terms such as *флотация* or *фильтрация* correspond in form, but they are indexed in different languages.

Table 2. The class-type relation links with multilingual descriptors.

Descriptor in Kazakh	Descriptor in English	Descriptor in Russian
Ұсату цехы	Crushing department	Дробильный цех
Ұсату бөлімшесі	Crushing workshop	Дробильная
Ұсату техникасы	Crushing unit	Дробильная техника
Ұсату машинасы	Crushing machine	Дробилка
Ұсату типі	Crushing Type	Тип дробления
Орташа ұсату	Medium crushing	Среднее дробление
Ұсақ ұсату	Fine crushing	Мелкое дробление

Notes: All the concepts concerning the term 'crushing' will be displayed with the equivalents in Kazakh, English, and Russian.

which the others are assigned as keywords, e.g., 'enrichment' = 'beneficiation.' Both terms have the same hierarchical category and meaning as in the Kazakh *кен байыту* and in the Russian *обогащение*.

Associative relations are established between the keywords that belong to the same or different semantic categories and arbitrary levels of the hierarchy, e.g., when searching for the descriptor 'metal production,' the AIS user and/or translator is offered to conduct additional searches for other descriptors (see Table 3).

Ontology in compiling multilingual thesaurus offers developed formalized means of describing the terms of the subject area that are used in modern intelligent information systems. Ontology consists of a set of concepts and statements, such as the classification of concepts, the relations between concepts, the hierarchy of concepts with respect to general-part and part-whole relations.

Thesauri in the modern interdisciplinary applied sciences such as computational linguistics and translation theory are considered to be linguistic resources, describing the relationships between lexical meanings of words in natural languages as a hierarchical system of synonymic groups, i.e., synsets (Fedotov et al., 2016). A synset, or synonymic set is a group of data elements that are considered semantically equivalent for the purposes of information retrieval. According to *WordNet*, a synset is a set of one or more synonyms that are interchangeable in some context. Each synset offers a definition and examples of the use of the word in a context. A term, word or phrase can appear in more than one synset and can have more than one category or part of speech. Each synset contains a list of synonyms or synonymic phrases and pointers describing the relations between other synsets. Various semantic relations interconnect these synsets:

1. Hyperonym: metal product → metal production; metal product → metallurgy
2. Hyponym: metal production → metal product; metallurgy → metal product
3. Has-member: metallurgy → metallurgist
4. Member-of: metallurgist → staff
5. Meronym: has-part: enrichment → flotation; enrichment → filtration
6. Antonym: magnetic separation → non-magnetic separation

In accordance with the definition of international standards, the Information Retrieval thesaurus is a normative dictionary that accurately indicates the relationship between terms and is intended to describe the content of documents (ISO, 1985, 1986, 2011, 2013). The controlled multilingual thesaurus is used for search queries to provide the translation of documents and user requests. Thus, the industry-specific terms of the multilingual thesaurus allow the AIS users and/or translators to find the optimal term, to use it as a search tool when searching for relevant information regardless of their location and health conditions.

When creating a controlled multilingual thesaurus, the first task is to select terms for thesaurus. There are several possible sources for terms when forming a multilingual thesaurus. Candidate terms for the controlled thesaurus are usually offered or checked by experts in the given field. Such experts who are responsible for language policy and representatives of state bodies, ministries and national departments have expertise on terminology. In addition, terms for thesauri may be derived from domain-specific texts using automated methods or manual processing of documents. When manually processing documents, indexers annotate the documents with the most relevant keywords, which are then reduced to a single list that may serve as the basis for the thesaurus. After the list of candidate terms is obtained, low-frequency terms are excluded from the list, since it is assumed that they are not informative enough to differentiate the separate documents. Relatively low-frequency terms can be removed from the list or presented as ascriptors of more general or frequent concepts.

As a result, our multilingual, controlled and interoperable thesaurus is composed of the following essential elements:

1. A list of terms systemized and classified according to subject areas and rubrics.
2. Term relations indicated by their hierarchical relative positions.
3. Equivalents of Kazakh terms in Russian and English.

Table 3. The associative relations with multilingual descriptors.

Descriptor in English	Descriptor in Kazakh	Descriptor in Russian
Metallurgy	Металлургия	Металлургия
Industry	Өндіріс	Промышленность
Raw materials	Шикізат	Сырье
Deposit	Кен орны	Месторождение
Metals	Металдар	Металлы
Processing methods	Өңдеу тәсілдері	Способы обработки
Equipment	Құрылғы	Оборудование
Transport	Көлік	Транспорт

Notes: All the concepts associated with the term 'metal production' will be displayed with the equivalents in Kazakh, English, and Russian.

Thus, in order to compile the up-to-date controlled multilingual thesaurus for industry-specific terms, it is necessary to provide the above-mentioned basic criteria.

4. Conclusion

Since Kazakhstan gained its sovereignty and independence in 1991, the elevation of the status of the Kazakh language has impacted the nation's education, economy, and industry. New generations of Kazakhstan's professionals in various subject areas must be fluent in Kazakh, Russian and English, and be mobile and effective in information retrieval. Right now, industry-specific translators have to make Kazakh–English translations through Russian as an intermediate language. Direct English–Kazakh translation occurs very rarely. Creating a new generation of dictionaries will enhance opportunities for Kazakhstan's translation services and strengthen the Kazakh language status as a state language.

Currently, IST is in demand for many areas of social and economic life of Kazakhstan. In accordance with the National Program of Industrial and Innovative Development of the Republic of Kazakhstan for 2015–2019 and spurred by the pandemic situation in 2019–2020, the controlled multilingual thesaurus, especially in covering terminology in the mining sector, gives equal opportunity to all AIS users and becomes the urgently necessary information retrieval tool. Such a thesaurus is needed because mining and metallurgical industry in Kazakhstan has a crucial role in providing sustainable development and ensuring the effective operation of subsoil use in areas such as exploration, development, production, processing and selling of solid minerals, as well as in improving the mineral resources potential of Kazakhstan with commercial and non-commercial reserves deposits.

Accordingly, the controlled multilingual and interoperable thesaurus of industry-specific terms is one of the effective tools for describing individual subject areas and is designed to promote communication and interaction between professionals, translators and AIS users for industry-specific areas irrespective of their location and health conditions. Unlike the multilingual dictionaries, the controlled multilingual thesaurus makes it possible to define the meaning of a term with the help of descriptions, correlations with other concepts and their groups and relate them to broader and narrower terms. Moreover, the content of the multilingual thesaurus can be used to feed the knowledge bases of artificial intelligence systems, as well as to build the foundation for a Kazakh language corpus by unifying and systemizing the Kazakh industry-specific terms in a multilingual way.

Acknowledgments

The article has been written within the Ministry of Education and Science of the Republic of Kazakhstan PhD Program (2017–2020), under grant agreement No. 01–57, dated 30 October 2017. The article has been written by a group of authors with the support of the State Material Reserve Committee of the Ministry of National Economy of the Republic of Kazakhstan to support developing industry-specific terminology with the involvement of experts from vulnerable communities. We thank the State Material Reserve Committee of the Ministry of National Economy of the Republic of Kazakhstan for its expertise in industry-specific terminology.

Conflict of Interests

The authors declare no conflict of interests.

References

Aitbaiuly, O. (2013). *Qazaq til bilimining termınologıasy maseleleri* [Terminological issues of Kazakh linguistics]. Almaty: Abzal-ai.

ANSI/NISO. (2003). *Information retrieval (Z39.50): Application service definition and protocol specification* (ANSI/NISO Z39.50-2003). Bethesda, MD: NISO Press. Retrieved from https://www.loc.gov/z3950/agency/Z39-50-2003.pdf

ANSI/NISO. (2005). *Guidelines for the construction, format and management of monolingual controlled vocabularies* (ANSI/NISO Z39.19-2005). Bethesda, MD: NISO Press. Retrieved from https://groups.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf

Bayekeyeva, A. (2018). Information technologies applied in compiling multilingual thesaurus. In *Conference Proceedings of the I International conference "Translation activity as interaction of cultures and languages" for students and young researchers* (pp. 325–331). Astana: L. N. Gumilyov Eurasian National University.

Bayekeyeva, A., Tazhibayeva, S., Shaheen A., Beisenova, S. Z., & Mamayeva, G. (2020). Multilingual thesaurus of industry-specific terms as major aids for translators. *Opción*, *36*(27), 1864–1900.

Beisenova, Z. (2011). *Otraslevaya terminologiya: Sistemnost, tipologiya, funkcionirovanie* [Industry-specific terminology: Consistency, typology, functioning] (2nd ed.). Astana: TsBO & MI LLP.

Beisenova, Z. (2014). *Intercultural approach of professional lexicon in veterinary medicine*. London: Aitmatov Academy.

Cabrė, M. T. (1998). *Terminology: Theory, methods and applications*. Amsterdam: John Benjamins Publishing Company.

Fedotov, A., Tussupov, J., Sambetbayeva, M., Fedotova, O., Sagnayeva, S., Bapanov, A., & Tazhibayeva, S. (2016). Classification model and morphological analysis in multilingual scientific and educational information systems. *Journal of Theoretical and Applied Information Technology*, *86*(1), 96–111.

IFLA. (2009). *Guidelines for multilingual thesauri* (IFLA

Professional Reports No. 115). Haia: International Federation of Library Associations and Institutions. Retrieved from http://www.ifap.ru/library/book411.pdf

International Organization for Standardization. (1985). *Documentation—Guidelines for the establishment and development of multilingual thesauri* (ISO 5964:1985). Geneva: International Organization for Standardization.

International Organization for Standardization. (1986). *Documentation—Guidelines for the establishment and development of monolingual thesauri* (ISO 2788:1986, 2nd ed.). Geneva: International Organization for Standardization.

International Organization for Standardization. (2011). *Information and documentation—Thesauri and interoperability with other vocabularies—Part 1: Thesauri for information retrieval* (ISO 25964–1:2011). Geneva: International Organization for Standardization.

International Organization for Standardization. (2013). *Information and documentation—Thesauri and interoperability with other vocabularies—Part 2: Interoperability with other vocabularies* (ISO 25964–2:2013). Geneva: International Organization for Standardization.

Kaidarov, A. (1993). *Qazaq termınologıiasyna jańasha kózqaras* [New view to Kazakh terminology]. Almaty: Rauan.

Kurmanbaiuly, S. (2014). *Qazaq termınologıiasy* [Kazakh terminology]. Almaty: Sardar.

Kydyralina, Z. (2014). Alimhan Ermekov i Kanysh Satpaev: Puti v nauke i linii sudeb. [Alimkhan Yermekov and Kanysh Satpayev: Ways in science and lines of fortune]. *Qazaqstan Tarihy*. Retrieved from https://e-history.kz/ru/library/show/24072

Library of Congress. (n.d.). Legislative indexing vocabulary. *Library of Congress*. Retrieved from https://id.loc.gov/vocabulary/subjectSchemes/liv.html

Marácz, L., & Adamo, S. (2017). Multilingualism and social inclusion. *Social Inclusion*, *5*(4), 1–4. http://doi.org/10.17645/si.v5i4.1286

Masterman, M. (1961). Semantic message detection for machine translation, using an interlingua. In *Proceedings of the 1961 international conference on machine translation of languages and applied language analysis* (Vol. 13, pp. 438–475). London: Her Majesty's Stationary Office. Retrieved from http://www.mt-archive.info/NPL-1961-TOC.htm

Nevskaya, I., & Tazhibayeva, S. (2015a). Sociolinguistic situation of Oguz Turks in Post-Soviet Kazakhstan. In T. Güdüz & M. Cengiz (Eds.), *Oguzlar. Dilleri, tarihleri ve kültürleri* [The Oguz. Their languages, histories and cultures] (pp. 321–335). Ankara: Hacettepe University.

Nevskaya, I., & Tazhibayeva, S. (2015b). Turkic languages of Kazakhstan: Problems and research perspectives. *Turkic Languages*, 18(1/2), 289–302.

Republic of Kazakhstan. (2020). Resolution No. 43 of the Chief State Sanitary Doctor of the Republic of Kazakhstan o*n further strengthening measures to prevent coronavirus infection among the population of the Republic of Kazakhstan*. Retrieved from https://online.zakon.kz/Document/?doc_id=37566604#pos=202;-54

Salton, G. (1979). *Dynamic information and library processing*. Upper Saddle River, NJ: Prentice Hall.

Shokin, Y., Fedotov, A., Zhizhimov, O., & Fedotova, O. (2015). Evolyuciya informacionnyh sistem: ot Web-saitov do sistem upravleniya informacionnymi resursami [Evolution of information systems: From web sites to information resource management systems]. *Bulletin of Novosibirsk State University. Series: Information Technology*, *13*(1), 117–134.

Shreider, Y. (1965). Ob odnoi modeli semanticheskoi teorii informacii [On one model of semantic information theory]. In A. Lyapunov (Ed.), *Problemy kibernetiki* [Problems of cybernetics] (Vol. 13, pp. 233–240). Moscow: Nauka.

Tussupov, J., Sambetbayeva, M., Fedotov, A., Fedotova, O., Sagnayeva, S., Bapanov, A., & Tazhibayeva, S. (2016). Classification model and morphological analysis in multilingual scientific and educational information systems. *Journal of Theoretical and Applied Information Technology*, *86*(1), 96–111.

Tussupov, J., Sambetbayeva, M., Fedotov, A., Sagnayeva, S., Bapanov, A., Nurgulzhanova, A., & Yerimbetova, A. (2016). Using the thesaurus to develop IT inquiry systems. *Journal of Theoretical and Applied Information Technology*, *86*(1), 44–61.

Well Drillers Association. (n.d.). About boreholes and wells. *Well Drillers Association*. Retrieved from https://www.welldrillers.org/faq/about-boreholes-and-wells

Yeskeldiyeva, B., & Tazhibayeva, S. (2015). Multilingualism in modern Kazakhstan: New challenges. *Asian Social Science*, *11*(6), 56–64.

About the Authors

Ainur Bayekeyeva is a PhD candidate in the Department of Translation Theory and Practice at L. N. Gumilyov Eurasian National University. Her main research interests are translation theory and practice, controlled thesaurus, industry-specific terminology, multimodal texts, multilingual Kazakh, Russian and English thesaurus. Within the framework of PhD program, granted by the Ministry of Education and Science of the Republic of Kazakhstan, she had internships on terminology and thesaurus at Goethe University Frankfurt (Germany) and Novosibirsk State University (Russian Federation).

Saule Tazhibayeva is a Professor, Doctor of Philology at the Philological Faculty, L. N. Gumilyov Eurasian National University. Her main research interests are general linguistics, Turkology, computer linguistics, Kazakh, Russian and English comparative and contrastive linguistics. The author of research articles and monographs on Turkic languages, cross subject and multidimensional studies published in Kazakhstan and abroad. She supervised the project "Turkic World of Kazakhstan: Language Variants, Cultural Archetypes and Self-Identification of Turkish Diaspora" granted by the Ministry of Education and Science of the Republic of Kazakhstan (2015–2017) and co-supervised the project "Interaction of Turkic Languages in the Post-Soviet Kazakhstan," jointly funded by Volkswagen Foundation (2014–2018).

Zhainagul Beisenova is a Professor, Doctor of Philology at the Philological Faculty, L. N. Gumilyov Eurasian National University. Her main research interests are industry-specific terminology, intercultural communication, comparative linguistics, and transcultural literature. Holder of such awards as *Best University Lecturer* and *Professor of the Year*, she is the author of research articles and monographs on industry-specific terminology published in Kazakhstan and abroad.

Aigul Shaheen is an Associate Professor, PhD at the Philological Faculty, L. N. Gumilyov Eurasian National University. Her main research interests are general linguistics, terminology, the Russian language and literature. She is the author of research articles and monographs on general linguistics and terminology published in Kazakhstan and abroad.

Aisaule Bayekeyeva is a Chief Civil Service expert at the State Material Reserve Committee of the Ministry of National Economy of the Republic of Kazakhstan. Her main research interests are translation theory and practice, the Kazakh language, language policy, language teaching and industry-specific terminology. She is responsible for implementing state language policy and state language use within the State Material Reserve Committee documents.