TECHNICAL SPECIFICATION

ISO/TS 24620-1

First edition
2015-03-15

Language resource management — Controlled natural language (CNL) —

Part 1:
Basic concepts and principles

Gestion des ressources linguistiques — Langage naturel contrôlé (CNL) —

Partie 1: Notions de base et principes

Reference number
ISO/TS 24620-1:2015(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web 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 documents 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).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

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

For an explanation on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the WTO principles in the Technical Barriers to Trade (TBT) see the following URL: Foreword - Supplementary information

The committee responsible for this document is ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 4, *Language resource management*.

ISO 24620 consists of the following parts, under the general title *Language resource management — Controlled natural language (CNL)*:

— *Part 1: Basic concepts and principles*

Introduction

The long history of the study of controlled natural language (CNL) has proved its effectiveness especially in technical documentation, technical communication, and business communication. In time, CNL has come to be applied in different ways in such fields as information management, librarianship, terminology management, and legal documents. Its commercial impact was established in the 1990s due to its effectiveness in information and communication technology applications like machine translation and mobile communication. Moreover, 'text simplification', a major task of CNL, promotes efficient communication with regard to all kinds of language use on the web. An example of this is the simplified English Wikipedia.

This part of ISO 24620 aims both to define major concepts related to CNL and to outline the scope of CNL and its various applications in relation to language resource management. These include the following:

a) the pre-editing of texts for CNL in preparation for machine translation;

b) the development of new or the re-use of existing controlled vocabularies for CNL;

c) the structuring and harmonization of content for content management;

d) technical writing including the formulation of standards;

e) facilitating communication with and for persons with disabilities (PWD), for instance, in ambient assisted living (AAL) or augmentative and alternative communication (AAC).

In this connection, special attention is given to aspects of interoperability: from technical interoperability through semantic interoperability to content interoperability. As a Technical Specification (TS), this part of ISO 24620 identifies the following:

— the environments of CNL used for the purposes and applications of all kinds (e.g. computer-assisted technical writing);

— the relationships to language resource management and related systems;

— the potential for new applications (e.g. in the processes, whether manual or automatic, of knowledge acquisition and knowledge fusion based on and linked to the web).

This part of ISO 24620 is the first in a planned series of International Standards on CNL. Subsequent parts will focus on issues specific to particular viewpoints and/or applications such as particular CNLs, CNL interfaces, the implementation of CNLs, and evaluation techniques for CNL.

Language resource management — Controlled natural language (CNL) —

Part 1: Basic concepts and principles

1 Scope

As part of a drive to provide international standards for language resource management, this part of ISO 24620 on controlled natural language (CNL) sets out the principles of CNL and its utilization together with the relevant supporting technology. However, this part of ISO 24620 also aims to introduce a general view of CNL with its objectives and characteristics and provide a scheme for classifying a range of CNLs. This part of ISO 24620 additionally specifies certain normalizing principles of CNLs that control the use of natural languages in particular domains and are also oriented towards areas of practical application. These areas include public administrative communications, search optimization, and the management of automatic question-answering systems, but the current version of this part of ISO 24620 does not address any issue involving these applications directly.

2 Terms and definitions

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

2.1
artificial language
language (2.11) that has been specifically devised for some applications

Note 1 to entry: The grammar of an artificial language is formulated systematically for some specific purposes of its used in practical applications especially in the area of human or human-machine communications.

2.2
authoring
writing a document such as a report, manual, article, or book

2.3
comprehension
understanding the *content* (2.4) of a document

2.4
content
information content
information contained in or conveyed by a *language* (2.11), which can be in a written, spoken, or some other forms such as images

2.5
content management
<language resource management> controlling the *content* (2.4) of a *text* (2.21) or the media in general while analysing or revising it

Note 1 to entry: This includes version control of revised documents, contents in versions of similar documents, and the management of relations between items in a document.

2.6
controlled natural language
controlled language
CNL
subset of *natural languages* (2.12) whose grammars and dictionaries have been restricted in order to reduce or eliminate both ambiguity and complexity

Note 1 to entry: As a generic, CNL is an uncountable noun that refers to the abstract properties of all controlled natural languages and not to a particular natural language or application for a specific purpose. It is engineered (i.e. constructed) with a view to reducing or eliminating ambiguity and complexity and aims both to make it easier for human readers [particularly non-native users, non-experts, and people with limited *comprehension* (2.3)] to read a *text* (2.21) and to improve the computational processing of a text.

Note 2 to entry: CNL is an engineered (i.e. constructed) language that is based on a particular natural language, but is more restrictive as regards lexicon, syntax, or semantics, while at the same time preserving most of its natural properties. Here, CNL is a countable noun.

2.7
controlled vocabulary
CV
list of lexical or phrasal items that are selected for the purpose of improving *readability* (2.15) in a particular domain

Note 1 to entry: Controlled vocabulary is also used in a more specific sense in applications such as

a) the field of information and documentation, where it is defined as a 'list of words or phrases authorized for indexing' [SOURCE: ISO 5127:2001(en), 4.2.2.1.03], and

b) in the field of health informatics, where it is defined as a 'finite set of values that represent the only allowed values for a data item' [SOURCE: CDISC Clinical Research Glossary version 8.0, 2009]. In the field of health informatics, these values may be codes, text, or numeric [SOURCE: ISO 11616:2012(en), 3.1.7].

Note 2 to entry: Most controlled vocabularies target a specific, narrow domain. Unlike CNL, they do not deal with grammatical issues (i.e. how to combine the terms needed to write complete sentences), but a good number of CNL approaches, especially domain-specific ones, include controlled vocabularies.

2.8
cooperative work
activity or result of working together to achieve the same goal

Note 1 to entry: Work carried out by more than one person in a collaborative way (e.g. technical writers and editors putting together a manual).

2.9
formal language
language (2.11) that has been devised for logical inferences or programming applications with a finite list of symbols and a finite set of formation rules based on these symbols that define well-formed sentences and also with a system that interprets these sentences

2.10
interoperability
<language resource management> achievement of partial or total compatibility between heterogeneous data models by the mapping of metadata

2.11
language
system of signs paired with meanings, thus, being used as a means of conveying information

2.12
natural language
NL
language (2.11) with its origin unknown, but continuously developing sometimes in idiosyncratic ways as is used conventionally for human communications

2.13
linguistic structure
composition of a *language* (2.11) at the level of sound, word, phrase, sentence, meaning, and discourse

Note 1 to entry: The science of language is understood to consist of phonology (sound), morphology (word units), syntax (sentential structure), semantics (meaning, information), and pragmatics (discourse, context).

2.14
pre-editing
modification of a *text* (2.21) before it is submitted to a specific processing (e.g. machine translation)

2.15
readability
ease of processing a *text* (2.21) for its *comprehension* (2.3)

2.16
re-use
use a document or data for purposes in addition to those for which it was originally designed

Note 1 to entry: Ability to use existing documents for new documents. This includes making a product manual for a new version of the product and one for a similar version.

2.17
rewriting
producing a new version of a *text* (2.21) by changing its lexical, sentential, or textual structures while keeping its original *content* (2.4)

2.18
simplification
process of reducing complexity

Note 1 to entry: A procedure such as *simplified language(2.19)* that makes *content* (2.4) simpler.

2.19
simplified language
language (2.11) generated through a *simplification* (2.18) process

2.20
special language
special-purpose language
SPL
language (2.11) used in a subject-specific field and also characterized by the use of specific linguistic means of expression

Note 1 to entry: The stricter the conventions of an SPL are systematized and made obligatory, the more they converge with CNL.

2.21
text
data in the form of characters, symbols, words, phrases, paragraphs, sentences, tables, or other character arrangements intended to convey a meaning and whose interpretation is essentially based on the knowledge of some *natural language* (2.12) or *artificial language* (2.1)

[SOURCE: ISO/IEC 2382-1:1993]

2.22
tractability
computational tractability
capability of being controlled, analysed, or generated

3 General view of controlled natural language

3.1 Overview

Controlled natural language (CNL) is a type of human language that is restricted for certain practical purposes such as machine translation and writing manuals. The main objective is to resolve lexical and structural ambiguity, and this is often achieved by reducing structural complexities, thereby, improving readability. However, the application of this restriction does not, in itself, mean that a controlled natural language (CNL) is a simplified language (SL).

CNLs can be further defined by

a) analysing other different kinds of controlled natural language,

b) comparing controlled natural languages with natural languages (NLs),

c) comparing controlled natural languages with special-purpose languages (SPLs), and

d) comparing controlled natural languages with artificial languages.

3.2 Properties of controlled natural languages

Over the years, CNL has been known as ‘processable’, ‘simplified’, ‘technical’, ‘structured’, or ‘basic’ language, but ‘controlled natural language’ is now the accepted term. A wide variety of such languages have been designed over the last four decades. They include Basic English,[10] Caterpillar Fundamental English (CFE),[12] Caterpillar Technical English (CTE), SBVR Structured English,[11] and Attempto Controlled English (ACE),[7]. They are used to improve translation and other communication between humans and provide natural and intuitive representations of formal notations.

NOTE SBVR stands for Semantics of Business Vocabulary and Rules.

Although many properties of CNLs and their environments have already been identified (see Reference [13]), CNL itself has the four properties listed below. Properties 1 and 2 below refer to the naturalness of CNL. On the other hand, properties 3 and 4 refer to the CNL control factor. The four properties of CNL are as follows:

a) it is based on one specific natural language (its ‘base language’);

b) the most important difference, but not necessarily the only difference, between the CNL and its base language is that the former is more restrictive with regard to lexicon, syntax, and semantics;

c) it is an engineered (i.e. constructed) language. This means that it is explicitly and consciously defined and is not the product of an implicit and natural process even though it is based in turn on a natural language that is the product of an implicit and natural process;

d) it preserves most of the natural properties of its base language and accordingly, speakers of the base language can intuitively and correctly understand texts in the controlled natural language, at least to a substantial degree.

These four properties describe application domains rather than the languages themselves. The PENS classification scheme[9] adds four dimensions: precision, expressiveness, naturalness and simplicity.

3.3 Controlled natural languages compared with natural languages

CNL is different from natural language and from artificial languages such as formal logics and programming languages. A CNL (e.g. controlled English) is a subset of the NL (in this case, English) that forms its base because every well-formed expression (word, phrase, or sentence) in the former is a well-formed expression in the latter. All CNLs have some traits of artificial language because they are intentionally and systematically modified or constrained by humans and the outcome involves rules and restricted vocabulary.

CNL is an application-oriented language. There can also be a range of particular CNLs. English, for example, has several dozen.

NL is used for communication between humans in general and contains ambiguous expressions. By contrast, CNL is a disambiguated language that either contains non-ambiguous expressions as primary text of CNL or needs additional disambiguating information for any ambiguous expressions in CNL text.

CNL is an expressively restricted language. The vocabulary and other types of expression of CNL are intended to be straightforward and transparent, and easy for non-native users, domain experts, and others, as the case may be, to understand.

3.4 Controlled natural languages compared with special-purpose languages

CNL resembles special-purpose language (SPL) in that it is an application-oriented language and its application areas mostly overlap with those of SPL, but it also has more general aims, more technical means of controlling each individual CNL, and different adequacy criteria. These will be addressed later on in this part of ISO 24620. The primary aim of CNL at an abstract meta-theoretical level is to resolve various types of ambiguity in text, thus, simplifying vocabulary and linguistic structures. CNL also has other, more specific aims with regard to specific domains. This issue is discussed in Clause 4. It follows that the overall design of CNL focuses on aspects of a language other than the specification of possible uses or purposes.

CNL is thus a modified version of NL. It is a language used by humans or computers for specific purposes and the modification takes the form of a variety of restrictions and constraints; one of these restrictions is simplification. Tools for generating CNL-based languages can be used to generate simplified languages like simplified English or simplified Japanese, for example, a Shakespeare play can turn into Shakespeare for Children or the elevated speech used by Japanese ladies in the Imperial Court can become a basis for the way that young ladies learning Japanese still speak to this day. These languages could still be mistaken for simplified languages generated by CNL-based systems and in this sense, SL is part of CNL.

CNL often has a restricted vocabulary constrained by terminological requirements. Since the term ‘cancer’ and ‘tumour’ are terms used in everyday language relating to health, more specific terms like ‘carcinoma’, ‘leukaemia’, and ‘lymphoma’, which refer to different types of cancer, might have to be used in medical diagnoses or prescriptions. In this sense, CNL for medicine can be said to be controlled. Spell-checkers and grammar checkers could also be seen as ways of controlling language and its use by imposing certain spelling preferences (e.g. ‘honor’ vs ‘honour’ or ‘convenor’ vs ‘convener’) or recommending the use of ‘that’ rather than ‘which’ or even ‘who’ for relative clause constructions. The outcome of these processes might not be simplification and it follows that not every CNL-based language is an SL.

Despite all the similarities and dissimilarities between individual CNLs and certain kinds of highly codified SPLs and SLs, they are CNL that retain the perspectives of the original motivations and designs for conceptualizing each of them. In practice, there can be no difference between them and the systems that process these languages can become interoperable. This interoperability is one of the aims of this part of ISO 24620.

3.5 Criteria of adequacy

The construction of CNL-based languages complies with the following three criteria of adequacy:

a) comprehension;

b) tractability;

c) operational generality.

The first criterion, comprehension, is a language-internal criterion that requires each CNL to ensure that various types of ambiguity are resolved and that the readability of text is improved. The other two are system-related. For the most part, the adequacy of tractability means computational tractability that guarantees human-machine interactions and other cooperative work involved in constructing CNL-based languages. The adequacy of operational generality leads to the interoperability of various systems, either SPL-based or SL-based, with varying designs and application purposes that generate various subtypes of NL or CNL.

NOTE Simplicity is not a criterion of adequacy for CNL-based languages or systems. In CNL, every language is considered to be complex as is every grammar system that parses and generates a language, although some aspects of a language or a system may be simpler than others. It follows that a simplified natural language is a kind of controlled natural language, but not every controlled natural language turns into a simplified language.

4 Benefits and usages of controlled natural language

4.1 Applications of controlled natural language

CNL will be developed for several target applications such as authoring, language learning, and human-machine interface or interaction. From the viewpoint of process, tasks will include the creation of documents in CNL, the editorial task of converting existing documents in NL to CNL, rewriting, and reproducing. For controlled authoring, human authors can cooperate with a machine to control human writing and produce CNL-expressed content. For educational applications, CNL can be used to control the processes and purposes of language learning, and for human-machine interaction, it can improve a variety of real-time online communications between humans and machines.

4.2 Users of controlled natural language

There are several categories of users who write and read or even speak and listen to CNL. The first class of users is that of humans. They include language-learners of all ages, writers who produce personal letters or technical documents, editors of newspapers and books, translators and interpreters, and they can be native users or non-native-users, and also people with cognitive or communication difficulties.

NOTE A CNL can be written and/or spoken, for example, spoken controlled English can be very effective if used for overseas broadcasting, especially when the listeners' first language is not English.

Intelligent machines can use CNL. The efficiency of equipment such as automatic translation machines, word processors, text-understanding processors, and systems for information retrieval and question answering can be substantially improved if they adopt an appropriate CNL.

4.3 Benefits of controlled natural language

CNL can be used to improve readability, resolve ambiguities, help accelerate reading, facilitate comprehension, and reduce the cost of all these processes. It is easier, for example, for a non-native user to understand CNL-based texts because they contain restricted vocabulary and tightly constrained syntactic structures. This also holds true for people who come from different domains or backgrounds because CNL disambiguates domain-specific terminology. CNL can reduce costs because it produces more controlled text and this makes it easier for humans and computer systems to translate it.

Another benefit is that written CNL texts such as the language resources employed in larger application scenarios (e.g. Semantic Web and decision-support systems) can be re-used. A number of aspects such as documents, texts, sentences, phrases, and terms can easily be retrieved and/or modified for re-use. This aspect is especially useful for CNL in industrial settings.