

ICS 01.040.01

CCS A 22



**NATIONAL STANDARD OF THE
PEOPLE'S REPUBLIC OF CHINA**

GB/T 20532-2025

Replacing GB/T 20532-2006

**Specification of part of speech (POS) tag of contemporary
Chinese for Chinese information processing (CIP)**

信息处理用现代汉语词类标记规范

Issued on: August 29, 2025

Implemented on: March 1, 2026

**Issued by: State Administration for Market Regulation;
Standardization Administration of the PRC**

Table of Contents

Preface	III
1 Scope	1
2 Normative References	1
3 Terms and Definitions	1
4 General Principles	1
5 Classification by word class and other segmentation units	2
6 Combined application of tag codes	6
7 Marking Code Table	7
8 Verification Method	10
Reference	11

Foreword

This document complies with the provisions of GB/T 1.1-2020 "Standardization Work Guidelines Part 1: Structure and Drafting Rules of Standardization Documents". Drafting.

This document supersedes GB/T 20532-2006 "Specification for Modern Chinese Part-of-Speech Tagging for Information Processing" and is consistent with GB/T 20532-2006.

Apart from structural and editorial changes, the main technical changes are as follows.

- a) The scope has been changed (see Chapter 1, Chapter 1 of the 2006 edition).
- (b) Regarding terminology and definitions, the term "Chinese information processing" has been changed to "???????" and the definition has been revised (see 3.1, 2006).
- (2.1 of the 2018 edition) added the term and definition of "Modern Chinese" (see 3.2); changed "Parts of Speech" to "Part of Speech (see 3.3, 2.3 in the 2006 edition).
- c) Regarding the basis for word class classification, the phrase "referring to the Outline of Grammar System for Secondary School Teaching" (see section 3.2 of the 2006 edition) has been deleted; the phrase "also" has been added.

The word class classification system was referenced from other language and writing standards, commonly used language reference books, and Chinese corpora. (See 4.2)

- d) For nouns, separate tags have been added for personal names in the noun subclass, labeled as nhf and nhs respectively (see 5.3.3); The examples of personal names in the noun subclass have been changed (see 5.3.3, 4.1.1.5 of the 2006 version); The examples

of organization names in the noun subclass have been changed.

Examples (see 5.3.5, 4.1.1.8 of the 2006 edition); Examples of changes to other proper nouns in the noun subclass (see 5.3.6, 2006 edition) (4.1.1.9); changed the noun subclass "time noun" to "time word" as the basic word class, tagged as t (see 5.4, 2006 edition).

4.1.1.2); The noun subclass "locative noun" has been changed to "locative word" as the basic word class, marked as l (see 5.5, 2006 edition).

4.1.1.3); The word "frontier" was deleted from the examples of locative nouns (see 4.1.1.4 in the 2006 edition); The word "frontier" was deleted from the noun subcategories.

Ethnic name (see 4.1.1.7 in the 2006 edition).

e) Regarding verbs, the examples of directional verbs have been revised (see 5.6.5, 4.1.2.5 of the 2006 edition); formal verbs have been added as verbs.

The word subclass is marked as vf (see 5.6.6); psychological verbs have been added as a verb subclass, marked as vs (see 5.6.7).

f) Regarding revisions to the numeral category, cardinal numbers have been added as a subcategory, marked as mc (see 5.9.1); ordinal numbers have been added as...

This is a subclass of numerals, denoted as mo (see 5.9.2).

g) Regarding pronouns, personal pronouns have been added as a subclass, marked as rp (see 5.11.1); demonstrative pronouns have also been added as a subclass.

The word subclass is marked as rd (see 5.11.2); interrogative pronouns have been added as a pronoun subclass, marked as ri (see 5.11.3).

h) Regarding the addition of new basic word classes, modal particles have been added as a basic word class, marked as y (see 5.16); revisions have been made to other segmentation units.

In terms of content, four categories of idioms-nominal idioms, verbal idioms, adjective idioms, and conjunctive idioms-have been removed.

Classes are tagged using a combination method (see 4.2.1 in the 2006 version); noun abbreviations, verb abbreviations, and adjective abbreviations have been removed.

Three subcategories of abbreviations are categorized by part of speech and are marked using a combination method (see 4.2.2 in the 2006 edition); noun morphemes have been removed.

The morphemes are categorized into three subcategories. characters, verbal morphemes, and adjective morphemes, and are marked using a combination method (see the 2006 edition). 4.2.5).

i) Regarding the newly added "Principles and Application of Tag Code Combination",

"Principles of Tag Code Combination" has been added (see 4.4); "Tag Code Combination" has been added.

The code combination application is explained in Chapter 6; syntax annotation for non-Chinese character strings has been added, and a combination method is used for annotation. Note (see 6.5).

j) Regarding other revisions, "other" has been changed to "other symbols" (see 5.19.7, 4.2.7 of the 2006 edition); "q" (quantifier) has been changed.

The code description for "quantity" is "according to common practice" (see Chapter 7, Chapter 5 of the 2006 edition); a "part of speech label" has been added.

The code table for recording segmentation units (see 7.1) has been added; a code table for other segmentation units (see 7.2) has been added; and a code table for combining part-of-speech tags has been added.

Table (see 7.3); "Verification Methods" (see Chapter 8) has been added.

This document was proposed by the Ministry of Education (Language).

This document is under the jurisdiction of the National Technical Committee on Standardization of Language and Writing (SAC/TC500).

This document was drafted by the Institute of Chinese Information Processing, Beijing Normal University, and the Institute of Applied Linguistics, Ministry of Education.

The main drafters of this document are: Yang Lijiao, Xiao Hang, Jin Guangjin, Fu Li, Liu Zhiying, Song Peiyan, Hu Renfen, Li Hongzheng, Li Hui, and Li Juan.

The release history of this document and the document it replaces is as follows.

- First published in 2006 as GB/T 20532-2006;
- This is the first revision.

1 Scope

This document specifies the labeling categories and marking codes for modern Chinese word classes and other segmentation units in information processing.

This document applies to Chinese information processing and language resource construction, modern Chinese teaching and research, international Chinese education, and Chinese dialects and archaism.

It can be used as a reference for Chinese information processing, etc., to support the development needs of intelligent information processing and the modernization of language governance capabilities.

2 Normative references

This document has no normative references.

3 Terms and Definitions

The following terms and definitions apply to this standard.

3.1 The process of application and other processing

Using computer technology and artificial intelligence methods, this method inputs, outputs, extracts, analyzes, understands, and generates information related to the pronunciation, form, and meaning of Chinese characters.

3.2 Modern Chinese (contemporary Chinese)

The common language of the Han nationality, which has been gradually established and widely used since the May Fourth Movement. 1)

3.3 segment unit

The basic unit used in Chinese information processing that has a defined grammatical function. 2)

- 1) Modern Chinese. This includes both spoken Chinese (represented by Standard Mandarin) and written Chinese (primarily modern vernacular). It is the current form of Chinese information processing. Primary language object.
- 2) Segmentation units. including words, phrases and other units as defined by the rules of this standard.

3.4 Part of speech; POS

The grammatical classification of words is mainly based on their grammatical function.

3.5 Tag

Code to label the categories of segmentation units in text.

4.1 Scope of the division unit

The segmentation units in this document include words, phrases, and other segmentation units such as idioms, abbreviations, prefixes, suffixes, and morphemes.