SAP BusinessObjects

Text Data Processing Language Reference Guide

- SAP BusinessObjects Data Services 4.1 Support Package 1 (14.1.1.0)

2012-12-20

SAP

Copyright

© 2012 SAP AG. All rights reserved.SAP, R/3, SAP NetWeaver, Duet, PartnerEdge, ByDesign, SAP BusinessObjects Explorer, StreamWork, SAP HANA and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG in Germany and other countries.Business Objects and the Business Objects logo, BusinessObjects, Crystal Reports, Crystal Decisions, Web Intelligence, Xcelsius, and other Business Objects products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of Business Objects Software Ltd. Business Objects is an SAP company.Sybase and Adaptive Server, iAnywhere, Sybase 365, SQL Anywhere, and other Sybase products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of Sybase, Inc. Sybase is an SAP company. Crossgate, m@gic EDDY, B2B 360°, B2B 360° Services are registered trademarks of Crossgate AG in Germany and other countries. Crossgate is an SAP company. All other product and service names mentioned are the trademarks of their respective companies. Data contained in this document serves informational purposes only. National product specifications may vary.These materials are subject to change without notice. These materials are provided by SAP AG and its affiliated companies ("SAP Group") for informational purposes only, without representation or warranty of any kind, and SAP Group shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP Group products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

2012-12-20

Contents

Chapter 1 Introduction....9

1.1 Overview of This Guide....9
1.1.1 About This Guide....9
1.1.2 Who Should Read This Guide....9

Chapter 2 Overview of Linguistic Analysis and Extraction....11

2.1 About Linguistic Analysis....11
2.2 About Extraction....11
2.2.1 About Customizing Extraction....12
2.3 Languages Modules Supported....13
2.4 Specialized Extraction Content....14

Chapter 3 Linguistic Analysis Support....15

3.1 Linguistic Analysis Language Feature Matrix....16
3.2 Segment Generation....19
3.3 Word Segmentation....19
3.3.1 White Space Languages....19
3.4 Case Normalization Rules....20
3.5 Stemming....21
3.5.1 Standard Inflectional Stemming....22
3.5.2 Expanded Inflectional Stemming....23
3.5.3 Inflectional Stemmer Guesser....23
3.5.4 Compound Word Stemming....24
3.5.5 Non-Decompounding Stemming....24
3.5.6 Derivational Stemming....24
3.5.7 Stemming Unknown Words....25
3.6 Part-of-Speech Support....25
3.6.1 Tag Name Conventions....26
3.6.2 Unfound Words....26
3.6.3 Tagged Stemming....26
3.6.4 Word Breaking....27

Chapter 4 **Extraction Support** **29**

4.1 Entity and Fact Extraction 29
4.1.1 Subtypes 30
4.2 Extraction Resource Files 30
4.3 Levels of Extraction Support for the Language Modules 31
4.4 Predefined Entity Type Support 33
4.4.1 Named Entities 33
4.4.2 Common Noun Mentions 41

Chapter 5 **Language Modules Reference** **43**

5.1 Catalan Language Reference 43
5.1.1 Linguistic Processing 43
5.1.2 Extraction 52
5.2 Chinese (Simplified) Language Reference 53
5.2.1 Linguistic Processing 53
5.2.2 Extraction 58
5.3 Chinese (Traditional) Language Reference 71
5.3.1 Linguistic Processing 72
5.3.2 Extraction 78
5.4 Croatian Language Reference 78
5.4.1 Linguistic Processing 78
5.4.2 Extraction 83
5.5 Czech Language Reference 83
5.5.1 Linguistic Processing 84
5.5.2 Extraction 89
5.6 Danish Language Reference 90
5.6.1 Linguistic Processing 90
5.6.2 Extraction 99
5.7 Dutch Language Reference 99
5.7.1 Linguistic Processing 100
5.7.2 Extraction 109
5.8 English Language Reference 109
5.8.1 Linguistic Processing 110
5.8.2 Extraction 119
5.9 French Language Reference 134
5.9.1 Linguistic Processing 134
5.9.2 Extraction 142
5.10 German Language Reference 154
5.10.1 Linguistic Processing 154

5.10.2 Extraction....168
5.11 Greek Language Reference....179
5.11.1 Linguistic Processing....179
5.11.2 Extraction....180
5.12 Hungarian Language Reference....180
5.12.1 Linguistic Processing....180
5.12.2 Extraction....182
5.13 Italian Language Reference....183
5.13.1 Linguistic Processing....183
5.13.2 Extraction....192
5.14 Japanese Language Reference....201
5.14.1 Linguistic Processing....201
5.14.2 Extraction....210
5.15 Korean Language Reference....216
5.15.1 Linguistic Processing....216
5.15.2 Extraction....224
5.16 Norwegian: Bokmål Language Reference....228
5.16.1 Linguistic Processing....228
5.16.2 Extraction....239
5.17 Norwegian: Nynorsk Language Reference....240
5.17.1 Linguistic Processing....240
5.17.2 Extraction....250
5.18 Polish Language Reference....251
5.18.1 Linguistic Processing....251
5.18.2 Extraction....252
5.19 Portuguese Language Reference....252
5.19.1 Linguistic Processing....252
5.19.2 Extraction....262
5.20 Romanian Language Reference....263
5.20.1 Linguistic Processing....263
5.20.2 Extraction....265
5.21 Russian Language Reference....265
5.21.1 Linguistic Processing....266
5.21.2 Extraction....271
5.22 Serbian Language Reference....275
5.22.1 Linguistic Processing....276
5.22.2 Extraction....281
5.23 Slovak Language Reference....282
5.23.1 Linguistic Processing....282
5.23.2 Extraction....288
5.24 Slovenian Language Reference....288

5.24.1 Linguistic Processing....289
5.24.2 Extraction....296
5.25 Spanish Language Reference....297
5.25.1 Linguistic Processing....297
5.25.2 Extraction....305
5.26 Swedish Language Reference....316
5.26.1 Linguistic Processing....316
5.26.2 Extraction....325
5.27 Thai Language Reference....325
5.27.1 Linguistic Processing....326
5.27.2 Extraction....327
5.28 Turkish Language Reference....327
5.28.1 Linguistic Processing....327
5.28.2 Extraction....329

Chapter 6 Voice of the Customer Content....331

6.1 Extracting Sentiments....333
6.1.1 English: Sentiment Extraction Examples....334
6.1.2 French: Sentiment Extraction Examples....335
6.1.3 German: Sentiment Extraction Examples....336
6.1.4 Spanish: Sentiment Extraction Examples....337
6.2 Extracting Emoticons....338
6.2.1 English: Emoticon Extraction Examples....338
6.3 Extracting Requests....339
6.3.1 English: Request Extraction Examples....340
6.3.2 French: Request Extraction Examples....340
6.3.3 German: Request Extraction Examples....341
6.3.4 Spanish: Request Extraction Examples....341
6.4 Extracting Profanities....342
6.4.1 English: Profanity Extraction Examples....342

Chapter 7 Enterprise Content....343

7.1 Extracting Membership Information....344
7.2 Extracting Management Change Events....346
7.3 Extracting Product Release Events....348
7.4 Extracting Merger Information....349
7.5 Extracting Organizational Information....350

Chapter 8 Public Sector Content....353

8.1 English: Types of Information Extracted....353

8.1.1 Public Sector Content Rule Sets–English....354
8.1.2 Extracting Action Events....355
8.1.3 Extracting Travel Events....363
8.1.4 Extracting Military Units....372
8.1.5 Extracting Organizational Information....373
8.1.6 Extracting a Person's Aliases....376
8.1.7 Extracting Information About a Person's Appearance....379
8.1.8 Extracting Information About a Person's Attributes....380
8.1.9 Extracting Information About a Person's Relationships....385
8.1.10 Extracting Spatial References....386
8.2 Simplified Chinese: Types of Information Extracted....387
8.2.1 Public Sector Entities–Simplified Chinese....387

Chapter 9 Common Mentions Content....393

9.1 Common Noun Mentions....393
9.1.1 COMMON_ADDRESS1 and COMMON_ADDRESS2....395
9.1.2 COMMON_CONTINENT....395
9.1.3 COMMON_COUNTRY....396
9.1.4 COMMON_FACILITY....396
9.1.5 COMMON_GEO_AREA....397
9.1.6 COMMON_GEO_FEATURE....397
9.1.7 COMMON_LOCALITY....398
9.1.8 COMMON_ORGANIZATION....398
9.1.9 COMMON_PERSON....399
9.1.10 COMMON_PRECURSOR....400
9.1.11 COMMON_REGION....400
9.1.12 COMMON_VEHICLE....400
9.1.13 COMMON_WEAPON....401

Index 403

Introduction

1.1 Overview of This Guide

Welcome to the *Language Reference Guide*.

SAP BusinessObjects Data Services text data processing software enables you to perform linguistic analysis of and extraction of content from unstructured text.

Linguistic analysis includes natural-language processing (NLP) capabilities, such as segmentation, stemming, and tagging, among other things. Extraction analyzes unstructured text, in multiple languages and from any text data source, and automatically identifies and extracts key entity types, including people, dates, places, organizations, or other information, from the text.

1.1.1 About This Guide

This guide contains two kinds of information:

- Overviews and conceptual information about the linguistic analysis and extraction features provided by the software.
- A reference section for each language supported by the software. It describes the behavior of the supported language modules during linguistic analysis and extraction.

1.1.2 Who Should Read This Guide

Users of this guide may need to enhance extraction in their text analytics application and should understand text data processing extraction concepts. However, users of this guide are not expected to understand or be familiar with the natural languages of the text being processed by the software. Similarly, users are not required to be familiar with linguistic principles. This document assumes the following:

- You are an application developer or consultant working on enhancing text data processing extraction.
- You understand your organization's text data processing extraction needs.

Overview of Linguistic Analysis and Extraction

The software includes language modules for the languages supported. Each language module consists of a set of files that include system dictionaries containing words to support the language processing operations for the given natural language. It is the language modules that enable linguistic analysis and extraction of unstructured text in a given language. Language modules use the following language processing technologies:

- Linguistic analysis to handle natural language processing
- Extraction to handle entity extraction

Related Topics

- Linguistic Analysis Support
- Extraction Support

2.1 About Linguistic Analysis

The software provides and uses sophisticated natural language processing capabilities for linguistic analysis of unstructured data. Some of these capabilities include:

- Segmentation–the separation of input text into its elements
- Stemming–the identification of word stems, or dictionary forms
- Tagging–the labeling of words' parts of speech

Related Topics

- Linguistic Analysis Support
- Language Modules Reference

2.2 About Extraction

Extraction is the process of discovering and presenting specific entities and facts that occur in unstructured text.

- Entities denote the names of people, places, things, dates, values, and so forth, that can be extracted from text. An entity is defined as a pairing of a standard form and its type. For example, `Winston Churchill/PERSON` is an entity in which `Winston Churchill` is the standard form and `PERSON` is the type.
- Facts are entities found during the extraction process, that represent relationships, events, sentiments, or requests. Facts are extracted based on extraction rules consisting of patterns that define the expressions to use to extract the information. The specialized voice of the customer content, for example, provides the rules that let you extract facts that represent sentiments and requests, as well as emoticons and profanities.

The language modules included with the software contain system dictionaries and provide an extensive set of predefined entity types. The extraction process can extract entities using these lists of specific entities. It can also discover new entities using linguistic models. Extraction classifies each extracted entity by entity type and presents this metadata in a standardized format.

Related Topics

- Extraction Support
- Predefined Entity Type Support
- About Customizing Extraction
- Languages Modules Supported
- Language Modules Reference
- Specialized Extraction Content

2.2.1 About Customizing Extraction

You can enhance the extraction process by creating and using:

- Dictionaries that contain information about entities. You can customize information about the entities your application must find.
- Extraction rules.

For details about enhancing extraction, refer to the *SAP BusinessObjects Data Services Text Data Processing Extraction Customization Guide*.

For certain language modules, you can also enhance extraction by using the specialized extraction content included in them.

Related Topics

- Specialized Extraction Content

2.3 Languages Modules Supported

The language modules provided by the software all support linguistic analysis. A subset of these language modules also support predefined entity extraction.

Language modules that support linguistic analysis and predefined entity extraction:

- Arabic
- Chinese (simplified)
- English
- Farsi
- French
- German
- Italian
- Japanese
- Korean
- Russian
- Spanish

Language modules that support linguistic analysis:

- Catalan
- Chinese (traditional)
- Croatian
- Czech
- Danish
- Dutch
- Norwegian Bokmål
- Norwegian Nynorsk
- Portuguese
- Serbian
- Slovak
- Slovenian
- Swedish

Language modules that support basic linguistic analysis:

- Greek
- Hebrew
- Hungarian
- Polish
- Romanian
- Thai
- Turkish

Note:
Not all linguistic analysis and extraction features are supported for all languages.

Related Topics

- Linguistic Analysis Language Feature Matrix
- Levels of Extraction Support for the Language Modules
- Language Modules Reference

2.4 Specialized Extraction Content

Certain language modules include specialized content that provides entity types and sets of rules that address specific needs:

Specialized Extraction Content	Description	Included in These Language Modules
Voice of the customer	Extracts specific information about your customers' needs (requests) and perceptions and problems (sentiments), as well as emoticons and profanities (English only)	English French German Spanish
Enterprise	Extracts enterprise-specific information, such as management changes and product releases	English
Public Sector	Extracts public-sector-specific information, such as events and relations	Arabic English Simplified Chinese
Common Mentions	Extracts common-noun variations of named entities	English

Related Topics

- Enterprise Content
- Public Sector Content
- Voice of the Customer Content

Linguistic Analysis Support

The software provides and uses these linguistic analysis features for multilingual natural language processing (NLP) of unstructured data:

Feature	Description
Language and encoding identification	The automatic recognition of the input language, for example, French or Japanese, and of various character encodings (such as Unicode UTF-8 and Code Page 1252).
Segment generation	The breaking of input text into segments of one or more complete paragraphs for more efficient processing.
Word segmentation	The separation of input text into its elements, such as words and punctuation.
Case normalization	The normalization of the initial letter of a word to upper or lower case. Used to counteract case changes related to document structure, such as title and heading capitalization.
Stemming	The identification of word stems, or dictionary forms, for text or single words.
Tagging	The labeling of words' parts of speech, for example, noun or verb.
Document analysis	The recognition of a document's major sections–paragraphs and sentences.
Tagged stemming	The identification of word stems for a word of a given part-of-speech.

Note:
Not all operations are supported for all languages.

Related Topics

- Linguistic Analysis Language Feature Matrix
- Segment Generation
- Word Segmentation
- Case Normalization Rules
- Stemming
- Part-of-Speech Support
- Tagged Stemming
- Language Modules Reference

3.1 Linguistic Analysis Language Feature Matrix

Linguistic analysis provides two levels of language support:

- Basic–Tagging is not supported
- Standard–Tagging is supported

The following table shows the status of each supported feature for each natural language.

Language	Multi-word Units	Word Seg-mentation	Compound Words	Inflection-al Stem-ming	Tagging	Tagged Stemming
Arabic	X	X		X	X	X
Catalan	X	X		X	X	X
Simplified Chi-nese		X	X*	X**	X	X
Traditional Chinese		X	X*	X**	X	X
Croatian	X	X		X	X	X
Czech	X	X		X	X	X

Language	Multi-word Units	Word Seg-mentation	Compound Words	Inflection-al Stem-ming	Tagging	Tagged Stemming
Danish	X	X		X	X	X
Dutch	X	X	X	X	X	X
English	X	X		X***	X	X
Farsi	X	X		X	X	X
French	X	X		X	X	X
Hebrew		X		X		
German	X	X	X	X	X	X
Greek	X	X		X		
Hungarian	X	X		X		
Italian	X	X		X	X	X
Japanese		X	X*	X	X	X
Korean		X	X*	X	X	X
Norwe gian:Bokmål	X	X		X	X	X
Norwe gian:Nynorsk	X	X		X	X	X
Polish	X	X		X		
Portuguese	X	X		X	X	X

Language	Multi-word Units	Word Seg-mentation	Compound Words	Inflection-al Stem-ming	Tagging	Tagged Stemming
Romanian	X	X		X		
Russian	X	X		X	X	X
Serbian	X	X		X	X	X
Slovak	X	X		X	X	X
Slovenian	X	X		X	X	X
Spanish	X	X		X	X	X
Swedish	X	X	X	X	X	X
Thai		X		X		
Turkish	X	X		X		

- * Compound analysis is supported by the expanded language module for the language.
- ** Because Chinese words are not inflected, the stems of all Chinese words are identical to their source forms. Therefore, stemming is not supported for Chinese.
- *** For English only, derivational stemming is also supported.

Related Topics

- Multiword Units
- Word Segmentation
- Stemming
- Compound Word Stemming
- Expanded Inflectional Stemming
- Derivational Stemming
- Part-of-Speech Support
- Tagged Stemming
- Language Modules Reference

3.2 Segment Generation

During the analysis of unstructured text, text processing objects operate on one segment of a data stream at a time. Segments are small units of text, including one or more complete paragraphs. Linguistic analysis operations break input streams into chunks. This chunking of the data stream is called segment generation.

Segment generation involves two steps: reading in the input text as a byte stream and breaking it into segments. The resulting segments contain associated metadata markup about the context text. These segments are then passed on for further linguistic analysis from which words, sentences, and paragraphs can then be extracted.

3.3 Word Segmentation

The word segmentation operation performs basic word breaking. It breaks text into the smallest, meaningful syntactic units, such as words or punctuation. The word segmenter also identifies idiomatic phrases, such as "case in point" or "out-of-the-box." These idiomatic phrases are processed as a single unit or word. Hyphenated words are not broken, since they are syntactic units. However, contractions (such as "don't") and elisions (such as "l'abri") are separated into their syntactic units.

3.3.1 White Space Languages

White space languages mark word boundaries with white space and punctuation marks. This group includes European, Balkan, and Middle Eastern languages, as well as Korean. Punctuation marks sometimes end a sentence, in which case they are used in sentence detection.

Non-white space languages include the Chinese languages, Japanese, and Thai (CCJT for short). Word segmentation in the CCJT languages occurs with a slightly different algorithm due to their structure. Because complete morphological analysis is required to perform word segmentation in these languages, the word segmentation, stemming, and part-of-speech tagging operations occur in a single step.

3.3.1.1 Multiword Units

By default, multiword units are segmented as a single unit, for example, "to and fro" and "Buenos Aires" are each segmented as one unit. However, you can turn this behavior off. In this case, multiword units are broken into their individual components. For example, "to and fro" is segmented into three units instead of one.

3.3.1.2 Punctuation

Word segmentors generally split off punctuation marks as separate units. This includes periods and commas, sentence-ending punctuation, and various quotation marks.

The following table summarizes punctuation-related segmentation conventions:

No Whitespace	If a punctuation mark is followed by a character and not by white space, it is not split off from its surrounding word. For example: "filename.filetype" is segmented as "filename.filetype".
Abbreviations	Abbreviations ending in a period are important exceptions to the general rule that splits punctuation from their terms; their periods remain with them.
Apostrophes	Contractions spelled with apostrophes (like can't, don't, etc. in English) are handled via language-specific rules.
Hyphens	Embedded and trailing hyphens are not split off from their words. Leading hyphens are not split off before a digit expression, for example, -1000 is segmented as one unit.

3.4 Case Normalization Rules

Case normalization provides case-normalized alternatives for words which, by their position in a sentence or because they occur in a title, may or may not appear with their inherent, meaningful capitalization. For instance, a proper noun like SAP is always capitalized, but a common noun like horse is only capitalized if it begins a sentence or occurs in a title. Therefore, if Horse is encountered, the case

normalizer provides the lower-case alternative so that later processing will not mistake Horse for a proper noun. The two resulting alternatives can then be passed on to the stemming or tagging operations.

Note:
Case normalization is not relevant to languages that do not distinguish between upper and lower case, for example, the CCJT languages, Arabic, Korean, Farsi, and Hebrew.

Case normalization depends on the type of sentence (normal sentence, title, or query) and the position of the word to be normalized in each sentence type. The important position to consider is the sentence-initial position, where special normalization rules may apply. Words directly following certain punctuation marks are also treated as if they are in sentence-initial position.

- Title sentence

  All capitalized words are normalized. For example, a newspaper heading would be normalized as:

  - `Cardinals Strike Out`(Cardinals | cardinals) (Strike | strike) (Out | out)

- Query sentence

  Lowercase words are normalized to their upper case variants. Capitalized and all-caps words are not normalized in query sentences.

  - `aaaa`: aaaa, Aaaa, AAAA
  - `aaaA`: aaaA, AaaA

- Normal sentence

  Capitalized words are normalized when they occur in sentence-initial position. All-caps words in sentence-initial position are also normalized. In other positions of normal sentences, capitalized and all-uppercase words are not normalized. For instance:

  - `Aaaa bbb Cccc`:(Aaaa | aaaa) (bbb) (Cccc)
  - `AAAA bbb CCCC`: (AAAA | Aaaa | aaaa) (bbb) (CCCC)

3.5 Stemming

Words like **speaks** or **speaking** have one stem– **speak**. Some words have more than one possible stem: **spoke**, for instance, may turn out, in context, to be the past tense of the verb **speak**, but it could also be the singular form of the noun **spoke**. A stem is a base form for one or more variant (source) forms found in text; it is the form referenced in the dictionary.

Stemming a word means finding and returning its stem. For example, rather than redundantly deal with **grind**, **grinds**, **grinding**, **ground**, and so on, all of these source forms can be recognized as variants of the single verb **grind**. **Ground** can also be a noun whose meaning is completely unrelated to the verb **grind**.

The example of indexing documents according to key words they contain can help to better understand the advantages of working with more abstract forms. If indexing is done naïvely, **grind**, **grinds**, **grinding**,

ground will be handled as unrelated words, and a query containing one of these variants will not return documents containing the other variants. With the use of a stemmer, however, all of the variants will be indexed under the base form grind (verb).

The stemmer the software uses receives input of a series of syntactic units (for example, **ground**) and associates each unit with one or more base forms (for example, **ground** , **grind**). The stemmer always returns all possible alternative stems for each input term.

The software distinguishes between standard inflectional stemming and derivational stemming. The stemmers are inflectional by default. Derived stemmers are indicated as such.

Inflectional stemming is provided for every supported language. At present, derivational stemming is supported only for English.

For some languages, two different inflectional stemmers are included–the standard inflectional stemmer and an expanded inflectional stemmer that is more permissive of variation in the input text.

The stemmers support several different variants of the stemming operation:

- The standard variant returns all possible normalized stems for the input. It also performs compound analysis in languages like German, such that compound words are broken into their component parts.
- The expanded variant covers the same normalization as the standard variant, but it is biased for recall by allowing wider variation in capitalization, accentuation, and similar features, as found in informal text.
- In German, the no-split stemmer supports compound stemming without breaking the compound into separate stems, which provides better browsability.
- In English, the derivational variant provides the root stem for morphologically derived words.

Related Topics

- Standard Inflectional Stemming
- Expanded Inflectional Stemming
- Derivational Stemming

3.5.1 Standard Inflectional Stemming

With inflectional stemming, words retain the part of speech (noun, verb, and so on) of the base forms. For example, the verb forms **speaks** and **speaking** remain verbs like the base form **speak**, even while incorporating changes related to person (first, second, third person), number (singular and plural), tense (present, past, future), aspect (progressive) or other grammatical features.

Here are some additional examples:

Example	Stems to
{aller, vais, vas, va, allons, allez, vont} [French]	**aller**
{reach, reaches, reached, reaching}	**reach**
{big, bigger, biggest}	**big**
{balloon, balloons}	**balloon**
{go, goes, going, gone, went}	**go**

The **bold** words are the stems (dictionary forms). The characters added to the stem (**es** in reach**es**, **s** in balloon**s**) are called inflections or affixes.

To handle unknown words such as neologisms, the standard stemmer contains a set of morphological rules that apply to words.

3.5.2 Expanded Inflectional Stemming

The expanded inflectional stemming dictionaries provide all the same functionality as the standard stemmers provided, and more. The expanded inflectional stemmer allows for certain non-standard word forms–for example, capitalization errors–as well as standard forms. Thus it can be used to process informal or imperfect text (such as email, online documents, or queries). The variation it handle includes case variation, hyphenation and unaccented characters among others. The expanded variant of the CCJT languages is designed for more granular stemming results suitable for index generation.

3.5.3 Inflectional Stemmer Guesser

The inflectional stemmer guesser contains morphological rules that can be applied to syntactic units that are unknown to the standard or expanded inflectional stemmer and, therefore, cannot be stemmed. The software provides inflectional stemmer guessers for English, French, German, and Spanish.

3.5.4 Compound Word Stemming

Compound words are those like bookmark or birdbath, formed by combining or concatenating several words. German is especially famous for its compounds, for example, **Bildungsroman** from **Bildung** "education" and **Roman** "novel", and **Weltanschauung** from **Welt** "world" and **Anschauung** "view".

The software performs compound analysis for German. In German, compounds are always separated into their component stems.

3.5.5 Non-Decompounding Stemming

The German language module includes a variant no-split stemmer that does not perform de-compounding in the stemmer. This stemmer stems the head of the compound, but does not split the compound into separate stems. For example, the plural compound **Bildungsromane** is stemmed to **Bildungsroman**, but is not split into component stems. The returned stem is always a single term; and since there is no compound boundary marker, the term cannot be broken up.

If alternate stems are possible, more than one stem may be returned, as with the standard and expanded stemmers.

3.5.6 Derivational Stemming

Derivational stemming involves cases in which words and stems may or may not have the same part of speech: a noun may be derived from a verb stem (as for **participation** and **participate**), or an adjective may be derived from a noun (as for **boyish** and **boy**). Here are more derivational examples:

- {introduction, introductory, introducer} from **introduce**
- {subcategory, categorize, categorization} from **category**
- {useful, usable, unusable} from **use**
- {reenlist} from **enlist**

Derivational stemming is currently supported for English only.

3.5.7 Stemming Unknown Words

The stemmer identifies the stems of all the standard words of a language. However, an unknown word, such as one not found in the system dictionary, will not have a stem. In general, the stemmer returns the input term as the stem itself. A complicating factor is that, due to case-normalization, the input to the stemmer may include more than one variant term for a given word. This means that one variant might be found while another might not be. By default, the stemmer returns the stems of found terms and removes unfound terms from the results.

For example, at the beginning of a sentence, the word Dogs would be normalized as the disjunction (Dogs | dogs). In such cases, the stemmer considers both members of the disjunction–both Dogs and dogs. Assume that lower-case dogs is in the stemmer dictionary, and that capitalized Dogs is absent. Since Dogs is not in the dictionary (and considered an unfound word), it would stem to Dogs itself. Since dogs is in the dictionary, it stems to dog. By default, the stemmer discards the unknown word Dogs and returns dog as the stem of the found variant. This is the default behavior.

If none of the case-normalized variants is found, then the stemmer returns all the case-normalized variants. For example, suppose the input sentence begins with the unknown word `Fbzzz`. The case normalizer returns the disjunction (`Fbzzz | fbzzz`). The stemmer finds neither one in the dictionary and returns both forms as stems.

Related Topics

- Case Normalization Rules

3.6 Part-of-Speech Support

The part-of-speech tagger identifies and labels the part of speech for each word in context. A word's part-of-speech is the grammatical category it falls into, such as noun or verb, along with subclass attributes of each of these major categories, such as singular or plural for nouns, and present or past tense for verbs.

For certain of its language modules, the software supports the use of two types of parts-of-speech tags. You can also use these tags when creating extraction rules:

- Umbrella tags–These tags identify major parts-of-speech at a high level, without breaking down the part of speech further than its overall function. For example, the `Nn` tag identifies all nouns, regardless of whether they are singular or plural, feminine or masculine, and so on.
- Complete tags–These tags identify the exact part-of-speech, along with its attributes. For example, the `Nn-Pl` tag identifies plural nouns, and `V-Pres-3-sg` identifies present tense, 3rd person singular verbs.

For specific details about the tag sets in each supported language, refer to the chapter for that language in the "*Language Modules Reference*" part of this guide.

3.6.1 Tag Name Conventions

Tags consist of feature names separated by hyphens. The first feature name is called a category tag. It usually specifies the high level part of speech of the word, for example, noun or verb, abbreviated as `Nn` and `V` respectively. When the tag contains more than one part-of-speech, as in `V/Adj` or `Det/Pron`, this indicates that the part-of-speech can be of either category.

Feature tags classify the word more precisely. They may indicate number (for example, plural and singular), person (for example, first, second or third), or tense (for example, present and past). Thus, the tag `V-Pres-3-Sg` indicates that the verb is present tense, third person singular.

When a feature appears in all lower case, as in the tag `Prep-para` from the Spanish tagger, it stands for a word in that language (here, Spanish para), and means that the word's distribution differs enough from that of other words of its category to rate its own feature. Such very specific features are listed in the language-specific tables.

For specific details about the tag sets in each supported language, refer to the chapter for that language in the "*Language Modules Reference*" part of this guide.

3.6.2 Unfound Words

Words not found in the tagger dictionary are passed to the relevant guesser to be assigned the most likely tag. The guesser assigns tags to unfound words based on a set of rules about the morphology of the given language. Capitalization information may also be used as capitalized words are also proper nouns in many languages. Combinations of alphabetic, numeric and optionally, punctuation characters tend to be guessed as proper nouns as well. Ordinal numbers are tagged either as noun or adjective, depending on the context. Internet and e-mail addresses are assigned the tag `Nn-Net`.

In the Asian languages, unfound words are assigned the tag `Nn` by default.

3.6.3 Tagged Stemming

The tagged stemming operation provides complete linguistic analysis of input text, including stemming with respect to part-of-speech information. This operation segments text into words and punctuation, performs document analysis, case normalization, and part-of-speech tagging. Then, given a term and

its part-of-speech tag, it performs stemming of the term. For example, for the input term-tag pair `children[Nn-Pl]`, the output is `child`.

3.6.4 Word Breaking

The word-breaking operation segments text into words and punctuation, performs document analysis, case normalization, and part-of-speech tagging.

Extraction Support

This section describes how extraction works when analyzing unstructured text.

4.1 Entity and Fact Extraction

Extracting entities from unstructured text tells us what the text is about–the people, organizations, places, and other parties described in the document. The extraction process involves processing and analyzing text, finding entities of interest, assigning them to the appropriate type, and presenting this metadata in a standard format.

The extraction process can extract entities using lists of specific named entities. It can also discover new entities using linguistic models.

Entities are often proper names, such as the names of specific and unique people, organizations, or places. Other specified entity types include currency amounts and dates, among others.

Each entity is defined as a pairing of a name and its type. For example:

- Canada/`COUNTRY`
- Pope John Paul/`PERSON`
- General Motors Corporation/`ORGANIZATION`/`COMMERCIAL`

Entity types play a crucial role in the definition of an entity. Entity types are used to classify entities extracted from documents and entities stored in a dictionary.

The extraction process presents this metadata in a standardized format, along with the entity's character offset and length in the document, and other attributes.

The software contains an extensive set of predefined entity types. You can optionally enhance the extraction process by using dictionaries and extraction rules.

For more details about creating dictionaries and extraction rules, refer to the *SAP BusinessObjects Data Services Text Data Processing Extraction Customization Guide*.

Related Topics

- Subtypes

4.1.1 Subtypes

A subtype indicates further classification of an entity type. It is a hierarchical specification that enables the distinction between different semantic varieties of the same entity type, such as commercial and educational organizations.

For example, SAP is an entity of type `ORGANIZATION` with a subtype `COMMERCIAL`, indicating a subcategory within the main category.

For those languages that support this features, their respective subtypes are described in the language's reference section in this guide.

Related Topics

- Entity and Fact Extraction

4.2 Extraction Resource Files

The extraction process uses several types of resource files: language modules, dictionaries, and extraction rule files. Some of these files are user-configurable, but not all.

This table provides a brief description of the resources that the extraction process uses:

Resource	Description
Language modules	A language module is a set of prepackaged, language-specific files, including dictionaries and other components that support a given operation in a given natural language. The dictionaries cover a large set of words for each supported language and are not user-configurable. Extraction relies upon the language modules to analyze text, extract entities and determine their type. For more information about specific language modules and their behavior, refer to their related chapter in the "*Language Modules Reference*" section of this guide.

Resource	Description
Dictionaries	Dictionaries are repositories of information about entities–their standard form and variant names, their entity types, and so on. Dictionaries are compiled into a proprietary format using the dictionary compiler tool.
Extraction rules	Extraction rule files contain linguistic and pattern-based rules that the software includes or that you can write using regular expression patterns to help you create links between entities, thereby extracting relation, event, and attributive-based facts. These rules are compiled using the extraction rule compiler.

For more information about writing and using extraction rules, refer to the *SAP BusinessObjects Data Services Text Data Processing Extraction Customization Guide*.

Related Topics

- Language Modules Reference

4.3 Levels of Extraction Support for the Language Modules

The language modules contain system dictionaries and configuration files required to perform entity extraction for several languages when analyzing text. All language modules include support for dictionaries and extraction rules.

Language modules are classified according to the level of linguistic analysis and extraction they support. They provide these levels of support:

- English–Of all the languages, English has the richest feature set. English supports a variety of predefined entity types, which also include predefined entity subtypes. It also supports parts-of-speech tags, the use of dictionaries and extraction rules, and the use of an advanced parsing capability for grammatical relations and pronominal co-reference resolution when processing extraction rules.
- Advanced–These languages support a variety of predefined entity types, dictionaries, and extraction rules. The advanced languages support extraction rule writing using syntactic units, the standard operators, the word stem and part-of-speech tag attributes to specify words, as well as a variety of linguistic construct markers such as noun phrases and clauses. The advanced languages are:
  - Arabic

    - Chinese: Simplified
    - Farsi
    - French
    - German
    - Italian
    - Japanese
    - Korean
    - Russian
    - Spanish
- Standard–These languages support noun phrase markers, dictionaries, and extraction rules. The standard languages support extraction rule writing using tokens, the standard operators, as well as the word stem and part-of-speech tag attributes to specify tokens. The standard languages are:
    - Catalan
    - Chinese: Traditional
    - Croatian
    - Czech
    - Danish
    - Dutch
    - Norwegian: Bokmål
    - Norwegian: Nynorsk
    - Portuguese
    - Serbian
    - Slovak
    - Slovenian
    - Swedish
- Basic–These languages support Linguistic Analysis features only such as multi-word tokens, word segmentation, or stemming. There is no noun-phrase support. The basic languages are:
    - Greek
    - Hebrew
    - Hungarian
    - Polish
    - Romanian
    - Thai

- Turkish

For more information about creating dictionaries and extraction rules, refer to the *SAP BusinessObjects Data ServicesText Data Processing Extraction Customization Guide*.

Related Topics

- Part-of-Speech Support

4.4 Predefined Entity Type Support

The entity type `NOUN_GROUP` is supported in all the language modules except in the basic language modules: Greek, Hebrew, Hungarian, Polish, Romanian, Thai, and Turkish. A `NOUN_GROUP` is any common noun sequence consisting of two or more related nouns, or modifier(s) plus noun(s), which are not identified as name, measure, or identifier.

4.4.1 Named Entities

The following table lists the predefined entity types in alphabetical order and indicates which languages support them.

Note:
For a list of additional public sector entities, see Public Sector Content.

Entity Type and Description	In Language Module:										
	ar*	zh (Simplified)	en	fa*	fr	de	it	ja	ko	ru	es
`ADDRESS1` Address	X	X	X	X	X	X	X	X			X

Entity Type and Description	In Language Module:										
	ar*	zh (Simplified)	en	fa*	fr	de	it	ja	ko	ru	es
ADDRESS2 Address (second part)			X				X				
CONTINENT Any of the continents	X	X	X		X	X	X	X			
COUNTRY Country name	X	X	X	X	X	X	X	X	X	X	X
CURRENCY Currency and currency expressions	X	X	X	X	X	X	X	X			X
DATE Date	X	X	X	X	X	X	X	X			X
DAY Day of the week	X	X	X	X	X	X	X	X			X
FACILITY Man-made structures	X	X	X						X		

Entity Type and Description	In Language Module:										
	ar*	zh (Simplified)	en	fa*	fr	de	it	ja	ko	ru	es
`GEO_AREA` Geographical area that is larger than a city and typically captures significant geographical areas		X	X		X	X	X	X	X	X	X
`GEO_FEATURE` Geographical name that does not fit in other `place/location` entity types	X	X	X	X	X	X	X	X	X	X	X
`HOLIDAY` Holidays and special days	X	X	X	X	X	X	X	X			X
`LANGUAGE` Noun referring to a language			X		X	X	X	X			X
`LOCALITY` City name	X	X	X	X	X	X	X	X	X	X	X

Entity Type and Description	In Language Module:										
	ar*	zh (Simplified)	en	fa*	fr	de	it	ja	ko	ru	es
MEASURE Measurement and measurement expressions	X	X	X	X	X	X	X	X			X
MISC_NUMERIC Number sequence followed by measure words		X		X							
MONTH Month, includes abbreviations	X	X	X	X	X	X	X	X			X
NAME_DESIGNATOR Designators that appear before a person's name, such as "c/o"			X				X	X			

Entity Type and Description	In Language Module:										
	ar*	zh (Simplified)	en	fa*	fr	de	it	ja	ko	ru	es
NIN National Identification Number. Social security number, including Canadian Social Insurance Numbers and French INSEE Numbers			X		X		X				
NOUN_GROUP Any common noun sequence consisting of two or more related nouns, or modifier(s) plus noun(s), which are not identified as name, measure, or identifier	X	X	X	X	X	X	X	X			X
ORGANIZATION Government, legal, or service agency including non-profit associations and institutions	X	X	X	X	X	X	X	X	X	X	X

Entity Type and Description	In Language Module:										
	ar*	zh (Simplified)	en	fa*	fr	de	it	ja	ko	ru	es
PEOPLE Name referring to a group of people based on country, ethnicity, or region	X	X	X		X	X	X	X			X
PERCENT Percents	X	X	X	X	X	X	X	X			X
PERSON Person's name	X	X	X	X	X	X	X	X	X	X	X
PHONE Phone numbers	X	X	X	X	X	X	X	X	X	X	X
PRODUCT Product name			X		X	X	X	X			X
PROP_MISC Any proper noun lacking an unambiguous type		X	X		X	X	X	X		X	X

Entity Type and Description	In Language Module:										
	ar*	zh (Simplified)	en	fa*	fr	de	it	ja	ko	ru	es
PUBLICATION Name of a newspaper, magazine, journal, and so on						X					
REGION Names of counties, prefectures, districts, and so on	X	X	X	X	X	X	X	X	X		X
SOCIAL_MEDIA Twitter handles and topics			X		X	X	X		X		X
TICKER Stock market ticker symbol			X		X		X				
TIME Time	X	X	X	X	X	X	X	X			X

Entity Type and Description	In Language Module:										
	ar*	zh (Simplified)	en	fa*	fr	de	it	ja	ko	ru	es
TIME_PERIOD Measures of time expressions	X	X	X	X	X	X	X	X			X
TITLE Title that is also used to refer to a person	X	X	X	X	X	X	X	X	X	X	X
URI Email address, URL, and so on	X	X	X	X	X	X	X	X	X	X	X
YEAR Year	X	X	X	X	X	X	X	X			X

Note:
* Information for the right to left languages entity names and extraction, namely Arabic, Farsi, and Hebrew is included in a separate supplement guide, *Right to Left Language Guide Reference Supplement*.

Related Topics

- Language Modules Reference
- Public Sector Entities–Simplified Chinese

4.4.2 Common Noun Mentions

The following table lists the predefined common noun mentions in alphabetical order and indicates which languages, other than English, support them. For English common noun mentions and conceptual information about common noun mentions, see Common Noun Mentions.

Note:
For a list of additional, public sector entities, see Public Sector Content.

Entity Type and Descrption	In Language Modules	
	Arabic	Simplified Chinese
`COMMON_ADDRESS1` Common names for addresses	X	
`COMMON_ADDRESS2` Common names for second part of addresses		
`COMMON_CONTINENT` Common names for continents	X	X
`COMMON_COUNTRY` Common names for countries including common nouns for geo-political entities for which the conventional labels do not apply, such as disputed territories or territories that have not been internationally recognized	X	X
`COMMON_FACILITY` Common names for man-made structures	X	X
`COMMON_GEO_AREA` Common names for geographical regions, districts, states, and provinces	X	X

Entity Type and Descrption	In Language Modules	
	Arabic	Simplified Chinese
COMMON_GEO_FEATURE Common names for places that are not geographical or political regions	X	X
COMMON_LOCALITY Common names for cities	X	X
COMMON_ORGANIZATION Common names for organizations	X	X
COMMON_PEOPLE Common names for people		X
COMMON_PERSON Common names for persons	X	X
COMMON_REGION Common names of counties, prefectures, districts, and so on	X	X

Related Topics

- Language Modules Reference
- Public Sector Entities–Simplified Chinese
- Common Mentions Content

Language Modules Reference

The Language Modules Reference provides a reference section for each language module supported by the software, and it includes the following information:

- The expected behavior of the language modules for all linguistic operations
- The predefined entity types supported by each language, with examples
- The umbrella and complete part-of-speech tags supported by each language, with examples

5.1 Catalan Language Reference

This chapter describes the behavior of the Catalan language module.

5.1.1 Linguistic Processing

This section describes the language-specific information on the processing of Catalan texts, including word segmentation, stemming, and tagging.

5.1.1.1 Character Encodings for Catalan

- iso_8859_1
- cp_1252
- utf_8, utf_16, ucs_4

5.1.1.2 Word Segmentation in Catalan

The Catalan segmenter follows all of the general segmentation rules in the white space languages. It has the following language-specific behavior.

Contracted forms, such as pel are not split. Elided pronouns preceding verbs are separated from the verb, as is the case with **m'** in **m'han vist**. Clitics appearing after the verb are split, both in their reduced (**'l** in **posa'l**) or full form (**-la** in **posa-la**).

Related Topics

- Word Segmentation
- White Space Languages

5.1.1.3 Stemming in Catalan

This section describes the standard stemmer and the expanded inflectional stemmer used for stemming in Catalan.

5.1.1.3.1 Standard Stemmer

The Catalan stemmer follows the general stemming rules as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. Proper nouns stem to themselves. Diminutive and superlative endings are removed from open classes. This is shown in the table below:

Category	Baseform	Examples
Noun	Non-diminutive masculine singular	vedelleta ->vedell
Proper Noun	Stemmed to themselves	Jordi -> Jordi
Verb	Infinitive	considerava -> considerar
Adverb	Source form	bé -> bé, activament -> activament

Category	Baseform	Examples
Pronoun	Masculine, nominative form	me -> jo

Catalan pronouns are stemmed in the following way. All uninflected forms stem to themselves. All personal pronouns maintain their number information. If applicable, these pronouns are stemmed to the nominative form. All other forms stem to the masculine, singular form. This is shown in the table below:

Text	Stem
tothom	tothom
elles	ells
em	jo
aquestes	aquest

Closed class words like determiners and ordinal numbers are stemmed to the masculine, singular form. Not-inflecting word categories stem to themselves, for example, conjunctions, cardinal numbers and prepositions:

Text	Stem
mitges	mig
ni	ni

Acronyms, abbreviations and multiword syntactic units stem to themselves:

Text	Stem
IVA	IVA

Text	Stem
tel.	tel.
davant de	davant de

Contracted forms are stemmed into their component part:

Text	Stem
pel	per=el
als	a=el

5.1.1.3.2 Expanded Inflectional Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text (such as email, online documents, or queries). See Expanded Inflectional Stemming for the general behavior. The specifics for Catalan follow.

Hyphenation

The expanded version accepts optional hyphenation for words for which normally have an obligatory hyphen.

Example	Output
Chupa-Chups	Chupa-Chups
ChupaChups	Chupa-Chups

Case Variants

The expanded version accepts lower case letters in addition to capital letters for those words where the capitals are obligatory. Initial capital letters will be lowercased to deal with those cases where common nouns are capitalized, for example, Amics de la Música.

Example	Output
IRPF	IRPF
irpf	IRPF
amics	amic
Amics	amic

Deaccented Characters

The expanded version accepts completely deaccented characters in addition to accented ones. It will also match words with acute (**é**, **ó**) accents to their grave counterpart (**è**, **ò**) and vice-versa. Additionally, for Catalan, the sequence **l.l** will be mapped to **l·l**.

Example	Output
nacio	nació
irlandés	irlandès
intel.ligent	intel·ligent

5.1.1.4 Part-of-Speech Tagging in Catalan

The following table shows the Catalan tag set available for using in defining custom entities. The tag names are accompanied by a brief description and one or more examples. The tag set makes no distinction for gender.

Umbrella Tag	Complete Tag	Description	Examples
`Abbr`	`Abbr`	Abbreviation	etc, Cia.

Umbrella Tag	Complete Tag	Description	Examples
Adj	Adj	Adjective, invariang for number	alfin, antidroga
	Adj-Ord-Pl	Plural, spelled out ordinal adjective	segons, primeres
	Adj-Ord-Sg	Singular, spelled out ordinal adjective	primer, segon
	Adj-Pl	Plural adjective	factibles
	Adj-Sg	Singular adjective	absurd, calent, capaç
Adv	Adv	Adverb	ací, abans, gairebé, fins
	Adv-Deg	Adverbs that can modify adjectives	bastant, força, gaire, massa, més, mig, molt, poc, prou, tan, tot
	Adv-Int	Interrogative adverbs	quan, on, a on, d'on, com, per què
	Adv-Rel	Adverbial relativizer	quan, com, on
Aux	Aux-Inf-be	Infinitive **ser**	ser
	Aux-Inf-have	Infinitive **haver**	haver
	Aux-anar	Auxiliary **anar**	vaig
	Aux-be	Auxiliary **ser**	serà
	Aux-have	Auxiliary **haver**	ha, han
Conj	Conj	Conjunction	si, perquè, mentre
	Conj-Coord	Coordinating conjunction	i, o, ni
	Conj-com	Conjuction **com**	com
	Conj-que	Conjunction **que**	que

Umbrella Tag	Complete Tag	Description	Examples
Det	Det-Def-Pl	Plural definite determiner	les, els
	Det-Def-Sg	Singular definite determiner	el, la, l'
	Det-Dem-Pl	Plural demonstrative determiner	aqueixes
	Det-Dem-Sg	Singular demonstrative determiner	aquest, això
	Det-Indef-Pl	Plural indefinite determiner or pronoun	gaires, molts, uns
	Det-Indef-Sg	Singular indefinite determiner or pronoun	bastant, gaire, quant, tant, molt, poc
	Det-Int-Pl	Plural interrogative determiner	quins
	Det-Int-Sg	Singular interrogative determiner	quin, quant
	Det-Poss-Pl	Plural possessive determiner	nostres, seues, llurs
	Det-Poss-Sg	Singular possessive determiner	teu, ma, llur
	Det-Rel-Sg	Singular relative determiner	qual
	Det-Rel-Pl	Plural relative determiner	quals
Interj	Interj	Interjection	ui!, eh?
Nn	Nn	Noun, invariable for number	atles, albatros, focus
	Nn-Net	URL or e-mail address	www.inxight.com info@inxight.com
	Nn-Pl	Plural noun	organitzacions, xarxes, casos, drets
	Nn-Sg	Singular noun	manera, exemple
Num	Num	Numeric expression, or cardinal number	2001, milions, dos
	Num-Ord	Ordinal number	1r, 2n, 3r, 4t, 5è

Umbrella Tag	Complete Tag	Description	Examples
Part	Part-Neg	The negation particle **no**	no
Prep	Prep	Preposition	amb, a causa de, darrera, en
	Prep-Det-a	Combination **a** and determiner	al, als
	Prep-Det-de	Combination **de** and determiner	del, dels, des del
	Prep-Det-per	Combination **per** and determiner	pel, pels
	Prep-a	Preposition **a**	a
	Prep-de	Preposition **de**	de, d'
	Prep-per	Preposition **per**	per
Pron	Pron	Pronoun	jo, tu, ell, això
	Pron-Adv	Adverbial pronoun	en, hi, n', -en
	Pron-Clitic	Clitic pronoun	s', 'ns, -hi
	Pron-Dem	Demonstrative pronoun	aquests
	Pron-Indef	Indefinite pronouns	moltes
	Pron-Int	Interrogative and exclamtive pronoun	qui, què, quant, quantes
	Pron-Oblq	Oblique pronoun	en, ho, ell, em
	Pron-Ord	Ordinal pronoun	tercer
	Pron-Poss	Possessive pronoun	el meu, la seva
	Pron-Rel	Relative pronoun	que, qui, què, qual
	Pron-es	**es** pronoun	es, se, s', -s
Prop	Prop	Proper noun or alpha numeric combination	Europa, FAO/OMS

Umbrella Tag	Complete Tag	Description	Examples
Punct	Punct	Other punctuation	: ; " ' { & /
	Punct-Close	Closed parenthesis	)
	Punct-Comma	Comma	,
	Punct-Open	Open parenthesis	(
	Punct-Sent	Sentence ending punctuation	. ! ?
V	V-Fin	Finite verb	reclamen, reconeix, passa, va
	V-Impv	Imperative verb	satisfacin, tracta
	V-Inf	Infinitive verb	arribar, mantenir, buscar
	V-PrPart	Present participle verb	creant, essent, donant
	V/Adj-PaPart-Pl	Plural past participle verb or adjective	elegits, encaminades
	V/Adj-PaPart-Sg	Singular past participle verb or adjective	fet, assenyalat, mancada

5.1.1.4.1 Unfound Words

Words not found in the tagger dictionary are passed to the Catalan guesser where they are assigned a tag based on a set of rules about Catalan morphology and capitalization. The following set of tagging rules are part of this module.

Verb tags are assigned according to the verb conjugation patterns. Internet and e-mail addresses are tagged as `Nn-Net`.

Words beginning with a capital letter or a number followed by a capital letter are guessed as proper nouns. Combinations of alphabetic, numeric and optionally, punctuation characters are also guessed as proper nouns. Combinations of digits and punctuation are tagged as numbers. A series of punctuation marks is tagged as punctuation.

5.1.1.5 Grouping in Catalan

A Catalan simple noun phrase is a noun with optional pre- and postmodifiers.

A premodifier is one or several adjectives, for instance:

- gran cilindrada
- nombrosos i ambiciosos projectes

A postmodifier may be an adjective, a noun or a prepositional phrase consisting of a form of **de** and another noun phrase.

- turisme responsable
- ciutat dormitori
- propostes de consultes populars
- multinacionals del Nord

Proper nouns are grouped just as common nouns are:

- antiga Iugoslàvia
- illes Balears
- Europa del nord
- Estatuts d' Autonomia

5.1.2 Extraction

This section describes the extraction-specific information for Catalan.

5.1.2.1 Predefined Entity Types

This section describes the predefined entity types supported by the Catalan language module and examples of each.

5.1.2.1.1 NOUN_GROUP

A Catalan simple noun phrase is a noun or series of nouns modified by optional pre- and postmodifiers. A premodifier can be an adjective or a series of coordinated adjectives. For example:

- pura casualitat
- nombrosos i ambiciosos projectes

A postmodifier can be an adjective or a prepositional phrase preceded by the preposition '**de**':

- gent impuntual
- eqip del ministeri

5.2 Chinese (Simplified) Language Reference

This chapter describes the behavior of the Simplified Chinese language module.

5.2.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Simplified Chinese texts, including word segmentation, stemming, and tagging.

5.2.1.1 Character Encodings for Simplified Chinese

- euc_cn
- gb_18030, gb_2312_80
- utf_8, utf_16, ucs_4

5.2.1.2 Word Segmentation in Chinese

The Chinese segmenter follows all of the general segmentation rules in the non-white space languages. It has the following language-specific behavior.

Bound morphemes like affixes are attached to content words. Also, classifiers are attached to preceding numbers. In the following Simplified Chinese example, 多 in 多媒体 is a prefix and 台 in 三台 is a classifier.

Text	Segmented
门市	门市
经营	经营
部门	部门
购得	购得
多媒体	多媒体
电脑	电脑
三台	三台

Hyphenated words are segmented into their separate parts. For instance:

Text	Segmented
北京 - 东京	北京
	-
	东京

Related Topics

- Word Segmentation
- White Space Languages

5.2.1.3 Stemming in Chinese

This section describes the standard stemmer and the expanded stemmer used for stemming in Chinese.

5.2.1.3.1 Standard Stemmer

Since Chinese words are not inflected, the stems of all words are identical to their source forms. This is true of the open class words listed in the following table as well as the closed class words.

Category	Baseform	Example
Noun	Source form	政府 -> 政府, 学生 -> 学生
Verb	Source form	负责 -> 负责, 保留 -> 保留
Adjective	Source form	小 -> 小, 必须 -> 必须
Adverb	Source form	非常 -> 非常

5.2.1.3.2 Expanded Stemmer

The expanded Chinese language modules provide more fine-grained segmentation and stemming results than the standard module. Its output is designed for optimized text indexing and search systems. The expanded module output differs from the standard stemmer in that classifiers are separated from numerals, prefixes and suffixes are separated from their head words, and compound analysis is performed.

Examples are shown below.

Classifiers are separated from numerals:

Text	Output
一本	一
	本

Prefixes and suffixes are separated from their head words:

Text	Output
女教师	女
	教师
小张	小
	张

Text	Output
发展部	发展
	部

Compounds are broken into their separate components:

Text	Output
布赖斯峡谷国家公园	布赖斯
	峡谷
	国家
	公园
彩色监定系统	彩色
	监定
	系统

The expanded variant supports all the same operations as the standard Chinese modules. However, its fine-grained output provides less contextual information for each term, and this ambiguity can compromise the accuracy of the tagging operations. For these operations, we recommend using the standard Chinese modules. The expanded variant is recommended for stemming purposes only.

5.2.1.4 Part-of-Speech Tagging in Chinese

The following table shows the Chinese tag set, which is the same for both Traditional and Simplified Chinese. The tag names are accompanied by a brief description and one or more examples. Simplified Chinese examples are given in GB encoding.

Umbrella Tag	Complete Tag	Description	Simplified Chinese Examples (GB)
`Adj`	`Adj`	Adjective	一流,大型

Umbrella Tag	Complete Tag	Description	Simplified Chinese Examples (GB)
Adv	Adv	Adverb	仅仅,非常
	Adv-BAN	Metaphor marker	般,似的
	Adv-Comp	Comparative adverb	最
	Adv-DENG	Post-nominal abbreviation	等
	Adv-Idiom	Idiomatic expression	寸草春晖,游人止步
Asp	Asp	Postverbal aspect marker	了,过,着
Aux	Aux	Auxiliary verb	应当,能
Cl	Cl	Classifier	张,副
Conj	Conj	Clausal conjoiner	不论,即使
	Conj-Nn	Noun conjoiner	及,和
Det	Det	Determiner	这,每,任何
Interj	Interj	Interjection	哇,喂
Nn	Nn	Common noun	东西,菜单,椅子
	Nn-Ascii	ASCII character noun	a, B
	Nn-Loc	locative noun	上,以内,之中
	Nn-Net	URL or email address	www.inxight.com
	Nn-Prop	Proper noun	香港,叶尔钦
	Nn-Time	Nominal time expression	今天, 周一, 上半年, 下午
Num	Num	Number	万,3, 5
Ord	Ord	Ordinal prefix	第
Part	Part	Sentence-final particle	吧, 吗

Umbrella Tag	Complete Tag	Description	Simplified Chinese Examples (GB)
Prep	Prep	Preposition	根据,以,由
	Prep-Assoc	Modification marker	的
	Prep-Assoc-ZHI	Noun-modification marker	之
	Prep-Assoc-DI	Verb-modification marker	地
	Prep-Assoc-DEI	Modification marker	得
Pron	Pron	Pronoun	她,我,你
Punct	Punct	Punctuation	..., –, ;, :
	Punct-Comma	Comma	,
	Punct-Open	Opening punctuation	（, {, 【
	Punct-Close	Closing punctuation	）, }, 】
	Punct-Sent	Sentence-ending punctuation	。
Quant	Quant	Quantifier	整个,众多
Verb	Verb	Verb	走,下雨,负责

5.2.2 Extraction

This section describes the extraction-specific information for Simplified Chinese.

5.2.2.1 Simplified Chinese Subtypes

Simplified Chinese supports subtypes in the types `FACILITY`, `GEO_AREA`, `GEO_FEATURE`, `ORGANIZATION`, `PEOPLE`, `REGION`, `URI`, `COMMON_FACILITY`, `COMMON_GEO_AREA`, `COMMON_GEO_FEATURE`, `COMMON_ORGANIZATION`, `COMMON_PEOPLE`, and `COMMON_PERSON`.

Related Topics

- Subtypes

5.2.2.2 Predefined Entity Types

This section describes the predefined entity types supported by the Simplified Chinese language module and examples of each. Click on the links to jump to that section: ADDRESS1, CONTINENT, COUNTRY, CURRENCY, DATE, DAY, FACILITY, GEO_AREA, GEO_FEATURE,HOLIDAY, LOCALITY, MEASURE, MISC_NUMERIC, MONTH, NOUN_GROUP, ORGANIZATION, PEOPLE, PERCENT, PERSON, PHONE, PROP_MISC, REGION, TIME, TIME_PERIOD, TITLE, URI, and YEAR.

Note:
The Simplified Chinese language module also extracts these public sector entities: `VEHICLE`, `WEAPON`, `COMMON_VEHICLE`, and `COMMON_WEAPON`.

For details about these public sector entities, refer to Public Sector Entities–Simplified Chinese.

5.2.2.2.1 ADDRESS1

Postal addresses:

- 北京市朝阳区建国门外大街甲12号新华保险大厦7层701室（100022）
- 上海市静安区南京西路1266号恒隆广场23楼2302-2304室（200041）
- 北京市朝阳区工体北路甲二号

5.2.2.2.2 CONTINENT

Any of the continents, for example:

- 亚洲
- 欧洲
- 南美洲

5.2.2.2.3 COUNTRY

Name of the countries and the names of geo-political entities for which the conventional labels do not apply. For example,

- 中国
- 美国
- 英国

- 巴勒斯坦
- 台湾

5.2.2.2.4 CURRENCY

Expressions denoting amounts of money:

- 33.8万元
- 港币五千万
- 一百四十四亿七千万美元

5.2.2.2.5 DATE

Dates are minimally composed of a number and month name:

- 7月2日
- 十月十七日

5.2.2.2.6 DAY

Names of the days of the week:

- 周一
- 周六

5.2.2.2.7 FACILITY

Man-made structures, extracted as one of the following subtypes:

- `AIRPORT`–The names of primarily man-made or man-maintained structures whose primary use is as transportation terminals. For example,
  - 首都国际机场
  - 浦东国际机场
  - 中正机场
- `BUILDGROUNDS`–The names of architectural and civil engineering structures, and outdoor spaces that are mainly man-made or man-maintained. There is no distinction with respect to their function, they could be civil or military facilities, they could be used for work or entertainment, or they could be monuments. For example,
  - 人民公园
  - 黄鹤楼
  - 克林姆林宫

- PATH–The names of primarily man-made or man-maintained structures that allows fluids, energy, persons, animals, or vehicles to pass from one location to another. For example,
  - 卢沟桥
  - 重庆南路
  - 王府井大街
- PLANT–The names of facilities composed by one or more buildings used for industrial purposes. For example,
  - 三峡工程
  - 切尔诺贝利核电站
  - 小浪底水库
- SUBAREA–The names of portions of facilities, typically architectural ones, that are able to contain people, animals, or objects. For Example,
  - 大雄宝殿
  - 椭圆形办公室

5.2.2.2.8 GEO_AREA

A geographical area larger than a city that captures a significant land mass, such as a continent or a group of countries, extracted as one of the following subtypes:

- DOMESTIC–The names of locations that do not cross national borders. For example:
  - 华南
  - 巴蜀
  - 杭嘉湖
- INTL–The names of locations that cross national borders. For example:
  - 大中华地区
  - 加勒比地区
  - 加沙地带

5.2.2.2.9 GEO_FEATURE

A place name extracted as one of the following subtypes:

- BOUNDARY–The names of locations such as borders. For example:
  - 南北回归线
  - 赤道
- CELESTIAL–The names of locations that are outside of the boundaries of the Earth. For example:

    - 地球
    - 冥王星
    - 北斗七星
- `LAND`–The names of locations that are geologically or ecosystemically designed, non-artificial locations. For example:
    - 峨眉山
    - 崇明岛
    - 珠江三角洲
- `WATER`–The names of locations that are bodies of water. For example:
    - 黄河
    - 长江
    - 西湖
    - 日月潭

5.2.2.2.10 HOLIDAY

Holidays and special days:

- 元宵节
- 中秋

5.2.2.2.11 LOCALITY

Name of a city:

- 北京
- 上海
- 苏州市

5.2.2.2.12 MEASURE

Measure expressions:

- 二百五十六公斤
- 5.5米

5.2.2.2.13 MISC_NUMERIC

Number sequence followed by measure words (not a major measure unit) or a noun:

- 八个
- 8000 多家

5.2.2.2.14 MONTH

Names of the months of the year:

- 6月份
- 八月

5.2.2.2.15 NOUN_GROUP

Noun groups can be simple or compound nouns with modifying adjectives:

- 新兴产业
- 高科技产品

5.2.2.2.16 ORGANIZATION

Government, legal, and service agencies, including non-profit organizations, fine arts groups, and other associations and institutions, extracted as one of the following subtypes:

- `COMMERCIAL`–The name of commercial organizations, such as major companies or corporations. For example:
  - 美洲银行
  - 花旗集团
  - 首创股份
  - 白云山制药股份有限公司

  Note:
  Any extracted `ORGANIZATION/COMMERCIAL` entities can be parsed and standardized using the Data Quality Data Cleanse transform by mapping them to one of the `FIRM` input fields.
- `EDUCATIONAL`–The names of institutions focused primarily in education. For example:
  - 交通大学
  - 清华
  - 浙大
- `ENTERTAINMENT`–The names of organizations focused primarily in entertainment. For examples:
  - 中央芭蕾舞团
  - 上海交响乐团
  - 月之海合唱团

- GOVERNMENT–The names of organizations related to government, politics, or the state. It also includes geopolitical entities that can function as political entities. For example:
  - 国务院
  - 海关总署
  - 水利部
  - 欧盟
  - 独联体
- MEDIA–The names of organizations focused on media, advertising, or publishing. For example,
  - 新华社
  - 时代周刊
  - 人民日报
- MEDICALSCIENCE–The names of organizations focused on medical care or research. For example:
  - 国家科学院
  - 中国科协
  - 中科院
- RELIGIOUS–the names of organizations focused on religion. For example:
  - 佛教
  - 基督教
  - 天主教
- SPORTS–The names of organizations focused on sports. For example:
  - 国家奥委会
  - 足球总会
  - 国际米兰俱乐部
- OTHER–Any organization that does not fit into a more specific subtype.
  - 中国共产党
  - 联合国
  - 全国总工会

5.2.2.2.17 PEOPLE

Groups of people extracted as the following subtype.

- NATIONALITY–Identifiable groups of people based on nationality.
- 中国人

- 美国人民

5.2.2.2.18 PERCENT

Percent expressions:

- 百分之五十
- 55.3%

5.2.2.2.19 PERSON

Variations of person names:

- 胡锦涛
- 毛 泽东
- 温家宝

5.2.2.2.20 PHONE

Phone numbers based on the Chinese format:

- 68316616

5.2.2.2.21 PROP_MISC

Any proper noun phrase that does not belong to one of the entity types specified by the other entities:

- 抗日战争
- 八国集团首脑会议
- 文化大革命

5.2.2.2.22 REGION

Different regions are extracted as one of the following subtypes:

- `MINOR-` The names of counties, prefectures, districts, or analogous geographical divisions or governmental units. For example,
  - 海淀区
  - 陆家嘴
  - 花莲县
- `MAJOR-` The major administrative divisions of countries, such as the provinces and territories of Canada, the administrative regions of France, and the states of the United States. For example,
  - 江苏省

- 新疆维吾尔族自治区
- 加利福尼亚

5.2.2.2.23 TIME

Clock times and time expressions:

- 8时
- 3点零5分

5.2.2.2.24 TIME_PERIOD

Measures of time duration:

- 两个月
- 1小时
- 五天

5.2.2.2.25 TITLE

Names of important positions in government, business, and other organizations:

- 主席
- 司法部长
- 总书记

5.2.2.2.26 URI

An address on the internet, extracted as one of the following subtypes:

- `EMAIL`–Email addresses, for example:
  - johndoe@businessobjects.com
  - support@inxight.com
- `IP`–IP addresses, for example:
  - 147.132.42.18
- `URL`–Internet addresses, for example:
  - www.businessobjects.com
  - http://www.google.com

5.2.2.2.27 YEAR

A year identifier and expressions based on years:

- 2005年
- 一九九四年

5.2.2.3 Common Noun Mentions

Common noun mentions refer to the use of common nouns to refer to entities such as organizations, persons, or facilities which would normally also be referred to by proper nouns.

This section describes the common noun mentions supported by the Simplified Chinese language module and examples of each. Click on the links to jump to that section:, COMMON_CONTINENT, COMMON_COUNTRY, COMMON_FACILITY, COMMON_GEO_AREA, COMMON_GEO_FEATURE, COMMON_LOCALITY, COMMON_ORGANIZATION, COMMON_PEOPLE, COMMON_PERSON, COMMON_REGION

5.2.2.3.1 COMMON_CONTINENT

Common nouns for the entirety of any continent:

- 大洲

5.2.2.3.2 COMMON_COUNTRY

Common nouns for countries as well as political regions:

- 王国
- 成员国
- 友邦
- 两岸三地

5.2.2.3.3 COMMON_FACILITY

Common nouns for man-made structures, extracted as one of the following subtypes:

- `AIRPORT`–Common nouns of primarily man-made or man-maintained structures whose primary use is as air transportation terminals. For example:
  - 机场
  - 空港
  - 候机大楼

- BUILDGROUNDS–Common nouns for architectural and civil engineering structures, and outdoor spaces that are mainly man-made or man-maintained. There is no distinction with respect to their function, they could be civil or military facilities, they could be used for work or entertainment, or they could be monuments. For example:
  - 大杂院
  - 建筑物
  - 停车场
- PATH–Common nouns for primarily man-made or man-maintained structures that allows fluids, energy, persons, animals, or vehicles to pass from one location to another. For example:
  - 高速铁路
  - 柏油路
  - 天桥
- PLANT–Common nouns for facilities composed by one or more buildings used for industrial purposes. For example:
  - 水电站
  - 厂矿
  - 水利枢纽
- SUBAREA–Common nouns for portions of facilities, typically architectural ones, that are able to contain people, animals, or objects. For Example:
  - 盥洗室
  - 卧房
  - 育婴房

5.2.2.3.4 COMMON_GEO_AREA

Common nouns for geographical regions, extracted as one of the following subtypes:

- DOMESTIC–Common nouns for locations that do not cross national borders:
  - 辖区
  - 国内
  - 非军事区
- INTL–Common nouns for locations that cross international borders:
  - 国际
  - 国内外

5.2.2.3.5 COMMON_GEO_FEATURE

Common nouns for places that are not geographical or political regions, extracted as one of the following subtypes:

- `BOUNDARY`–Common nouns for locations such as a border:
  - 国界
  - 边境线
- `CELESTIAL`–Common nouns for locations outside of Earth:
  - 小行星
  - 星系
  - 星球
- `LAND`–Common nouns for geologically or ecosystemically designed non-artificial locations:
  - 平原
  - 群岛
  - 戈壁
- `WATER`–Common nouns for bodies of water:
  - 江
  - 河
  - 湖
  - 海

5.2.2.3.6 COMMON_LOCALITY

Common nouns for cities:

- 全市
- 小镇
- 省会

5.2.2.3.7 COMMON_ORGANIZATION

Common nouns for organizations, extracted as one of the following subtypes:

- `COMMERCIAL`–Common nouns for companies:
  - 公司
  - 集团
  - 财团

  - 银行
- `EDUCATIONAL`–Common nouns for institutions focused on education:
  - 学院
  - 高校
  - 母校
- `ENTERTAINMENT`–Common nouns for institutions focused on entertainment:
  - 弦乐队
  - 马戏团
  - 文工团
- `GOVERNMENT`–Common nouns for institutions related to government, politics, or the state:
  - 军队
  - 机关
  - 法院
- `MEDIA`–Common nouns for institutions related to the media:
  - 传媒
  - 电视台
  - 报社
- `MEDICALSCIENCE`–Common nouns for institutions related to medical science:
  - 研究所
  - 综合医院
- `OTHER`–Common nouns for organizations that do not fit into a more specific subtype:
  - 协会
  - 理事会
  - 联合会
- `RELIGIOUS`–Common nouns for institutions related to religion:
  - 教宗
  - 主教团
  - 教会组织
- `SPORTS`–Common nouns for institutions related to sports:
  - 蓝球队
  - 羽毛球队

- 运动联合会

5.2.2.3.8 COMMON_PEOPLE

Common nouns for peoples, extracted as the following subtypes:

- `NATIONALITY`–Nationalities without modifiers:
  - 人民

5.2.2.3.9 COMMON_PERSON

Common nouns for persons, extracted as one of the following subtypes:

- `GROUP`–Common nouns for groups of persons:
  - 股民
  - 小两口
  - 中青年
- `INDIVIDUAL`–Common nouns for individual persons:
  - 老大爷
  - 师父
  - 导演

5.2.2.3.10 COMMON_REGION

Common nouns for different regions, extracted as one of the following subtypes:

- `MAJOR`–Common nouns for major administrative divisions of countries. For example,:
  - 省份
  - 自治区
- `MINOR`–Common nouns for the entirety of district areas. For example,:
  - 郡
  - 县
  - 区

5.3 Chinese (Traditional) Language Reference

This chapter describes the behavior of the Traditional Chinese language module.

5.3.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Traditional Chinese texts, including word segmentation, stemming, and tagging.

Note:
Because the linguistic processing algorithms used for Traditional and Simplified Chinese are the same, all the examples in this section are Simplified Chinese unless otherwise mentioned. For Traditional Chinese, only the character encoding would differ.

5.3.1.1 Character Encodings for Traditional Chinese

- big5
- utf_8, utf_16, ucs_4

5.3.1.2 Word Segmentation in Chinese

The Chinese segmenter follows all of the general segmentation rules in the non-white space languages. It has the following language-specific behavior.

Bound morphemes like affixes are attached to content words. Also, classifiers are attached to preceding numbers. In the following Simplified Chinese example, 多 in 多媒体 is a prefix and 台 in 三台 is a classifier.

Text	Segmented
门市	门市
经营	经营
部门	部门

Text	Segmented
购得	购得
多媒体	多媒体
电脑	电脑
三台	三台

Hyphenated words are segmented into their separate parts. For instance:

Text	Segmented
北京 - 东京	北京
	-
	东京

Related Topics

- Word Segmentation
- White Space Languages

5.3.1.3 Stemming in Chinese

This section describes the standard stemmer and the expanded stemmer used for stemming in Chinese.

5.3.1.3.1 Standard Stemmer

Since Chinese words are not inflected, the stems of all words are identical to their source forms. This is true of the open class words listed in the following table as well as the closed class words.

Category	Baseform	Example
Noun	Source form	政府 -> 政府, 学生 -> 学生
Verb	Source form	负责 -> 负责, 保留 -> 保留
Adjective	Source form	小 -> 小, 必须 -> 必须
Adverb	Source form	非常 -> 非常

5.3.1.3.2 Expanded Stemmer

The expanded Chinese language modules provide more fine-grained segmentation and stemming results than the standard module. Its output is designed for optimized text indexing and search systems. The expanded module output differs from the standard stemmer in that classifiers are separated from numerals, prefixes and suffixes are separated from their head words, and compound analysis is performed.

Examples are shown below.

Classifiers are separated from numerals:

Text	Output
一本	一
	本

Prefixes and suffixes are separated from their head words:

Text	Output
女教师	女
	教师

Text	Output
小张	小
	张
发展部	发展
	部

Compounds are broken into their separate components:

Text	Output
布赖斯峡谷国家公园	布赖斯
	峡谷
	国家
	公园
彩色监定系统	彩色
	监定
	系统

The expanded variant supports all the same operations as the standard Chinese modules. However, its fine-grained output provides less contextual information for each term, and this ambiguity can compromise the accuracy of the tagging. For these operations, we recommend using the standard Chinese modules. The expanded variant is recommended for stemming purposes only.

5.3.1.4 Part-of-Speech Tagging in Chinese

The following table shows the Chinese tag set, which is the same for both Traditional and Simplified Chinese. The tag names are accompanied by a brief description and one or more examples. Traditional Chinese examples are given in the Big5 encoding.

Umbrella Tag	Complete Tag	Description	Traditional Chinese Examples (Big5)
Adj	Adj	Non-predicative adjective	一流,大型
Adv	Adv	Adverb	僅僅,非常
	Adv-BAN	Metaphor marker	般,似的
	Adv-Comp	Comparative adverb	最
	Adv-DENG	Post-nominal abbreviation	等
	Adv-Idiom	Idiomatic expression	寸草春暉,游人止步
Asp	Asp	Postverbal aspect marker	了,過,著
Aux	Aux	Auxiliary verb	應當,能
Cl	Cl	Classifier	張,副
Conj	Conj	Clausal conjoiner	不論,即使
	Conj-Nn	Noun conjoiner	及,和
Det	Det	Determiner	這,每,任何
Interj	Interj	Interjection	哇,喂

Umbrella Tag	Complete Tag	Description	Traditional Chinese Examples (Big5)
Nn	Nn	Common noun	東西,菜單,椅子
	Nn-Ascii	ASCII character noun	a, B
	Nn-Loc	Locative noun	上,以內,之中
	Nn-Net	URL or email address	www.inxight.com
	Nn-Prop	Proper noun	香港,葉爾欽
	Nn-Time	Nominal time expression	今天, 周一, 上半年 下午
Num	Num	Number	萬, 3, 5
Ord	Ord	Ordinal number	第
Part	Part	Sentence-final particle	吧, 矣
Prep	Prep	Preposition	根據,以,由
	Prep-Assoc	Modification marker	的
	Prep-Assoc-ZHI	Noun-modification marker	之
	Prep-Assoc-DI	Verb-modification marker	地
	Prep-Assoc-DEI	Modification marker	得
Pron	Pron	Pronoun	她,我,你
Punct	Punct	Punctuation	..., –, ;, :
	Punct-Comma	Comma	,
	Punct-Open	Opening punctuation	（, ｛, 【
	Punct-Close	Closing punctuation	）, ｝, 】
	Punct-Sent	Sentence-ending punctuation	。
Quant	Quant	Quantifier	整個,眾多
Verb	Verb	Verb	走,下雨,負責

5.3.2 Extraction

This section describes the extraction-specific information for Traditional Chinese.

5.3.2.1 Predefined Entity Types

This section describes the predefined entity types supported by the Traditional Chinese language module and examples.

5.3.2.1.1 NOUN_GROUP

Chinese noun groups consist minimally of one noun, but may include more than one noun, as in:

- 主細胞
- 國際勞工組織

5.4 Croatian Language Reference

This chapter describes the behavior of the Croatian language module.

5.4.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Croatian texts, including word segmentation stemming, and tagging.

5.4.1.1 Character Encodings for Croatian

- iso_8859_2

- cp_1250
- utf_8, utf_16, ucs_4

5.4.1.2 Stemming in Croatian

This section describes the standard stemmer and the expanded stemmer used for stemming in Croatian.

5.4.1.2.1 Standard Stemmer

The standard Croatian stemmer follows the general stemming rules as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Examples
Noun	vojske -> vojska , ljudi -> èovjek , mjesta -> mjesto
Verb	udovoljava -> udovoljavati , ponude -> ponuditi , komentirao -> komentirati
Adjective	srbijansku -> srbijanski , spremni -> spreman , izborni -> izboran
Adverb	kako -> kako , sada -> sada , opet -> opet

5.4.1.2.2 Expanded Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text, such as email, online documents, or queries. In Croatian, this includes accented characters missing their diacritics and proper names without word-initial capitalization.

Case Variants

The expanded version accepts lower case letters in addition to capital letters for those words where the capitals are obligatory. Initial capital letters will be lowercased to deal with those cases where common nouns are capitalized, for example, Splet Svjetskih Mreža.

Example	Output
Hrvata	Hrvat
hrvata	Hrvat
Plovka	plovka
plovka	plovka
Splet	splet
splet	splet

Deaccented Characters

The expanded version also allows deaccented characters in place of accented ones.

Example	Output
bajaèica	bajaèica
bajacica	bajaèica

5.4.1.3 Part-of-Speech Tagging in Croatian

The following table shows the Croatian tag set available for using in defining custom entities. The tag names are accompanied by a brief description and one or more examples. The tag set makes no distinction for gender.

Umbrella Tag	Complete Tag	Description	Examples
`Abbr`	`Abbr`	Abbreviation	HDZ, RH

Umbrella Tag	Complete Tag	Description	Examples
Adj	Adj	Adjective absolutive	dobar, velik
	Adj-Comp	Adjective comparative	bolji, veći
	Adj-Sup	Adjective superlative	najbolji, najveći
Adv	Adv	Adverb absolutive	brzo, mnogo
	Adv-Comp	Adverb comparative	brže, više
	Adv-Sup	Adverb superlative	najbrže, najviše
Conj	Conj-Co	Conjunction coordinate	a, i
	Conj-Sub	Conjunction subordinate	jer, da
Enum	Enum	Enumeration	etc.
Interj	Interj	Interjection	hej, jao
Nn	Nn-Acc	Noun accusative	žene, profesori
	Nn-Case	Noun case other than nominative and accusative	ženama, profesorom
	Nn-Nom	Noun nominative	žene, profesor
Num	Num	Numeral	dvanaest, sedamdeset
	Num-Acc	Numeral accusative	jednu, jednog
	Num-Card	Numeral cardinal	tri, četiri
	Num-Case	Numeral case other than nominative and accusative	jednom, dvama
	Num-Nom	Numeral nominative	jedan, dva
	Num-Ord	Numeral ordinal	prvi, drugi
Prep	Prep	Preposition	za, na, u

Umbrella Tag	Complete Tag	Description	Examples
Pron	Pron	Pronoun	obje, vas, ovi, moji, koja
	Pron-Pers	Pronoun personal	ja, ti
	Pron-Poss	Pronoun possessive	tvoji, naši
	Pron-Ref	Pronoun reflexive	se
Prop	Prop	Proper name	Zagreb
Punct	Punct	Punctuation	.:;-
	Punct-Close	Punctuation mark closing	)
	Punct-Comma	Punctuation mark comma	,
	Punct-Open	Punctuation mark opening	(
	Punct-Sent	Punctuation mark sentence	.!?
V	V-Aux-Clit	Verb auxiliary clitic	je, sam
	V-Fin	Verb finite	radimo, nose, nosi
	V-Inf	Verb infinitive	raditi, nosi
	V-Part	Verb participle	misleći, uzimajući, nosili, nosio

5.4.1.4 Grouping in Croatian

A simple noun phrase in Croatian consists of a noun or nominal pronoun with optional adjectival specifiers and optional nominal and prepositional complements: velika soba, soba na katu, nepoznat netko.

The specifier is an adjective, adjectival pronoun, adjectival numeral or series of adjectives, adjectival pronouns, and/or adjectival numerals.

- veliki stol, moj stol, moj prvi stol, moj veliki stol

A complement can include nouns, or a preposition plus a (possibly modified) noun combination.

- buka motora
- rad na crno

All prepostions are allowed as PP complements:

- put u Pariz, stepenice na terasu, pasta za zube

5.4.2 Extraction

This section describes the extraction-specific information for Croatian.

5.4.2.1 Predefined Entity Types

This section describes the predefined entity types supported by the Croatian language module and examples of each.

5.4.2.1.1 NOUN_GROUP

Croatian noun groups can consist of one or more nouns with optional premodifiers.

The premodifiers can consist of zero or more adverbs followed by one or more (possibly coordinated) adjectives, adjectival pronouns, and/or adjectival numerals.

For example:

- milijardi kuna
- skoroj privatizaciji
- minimalno potrebnog broja

5.5 Czech Language Reference

This chapter describes the behavior of the Czech language module.

5.5.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Czech texts, including word segmentation and stemming.

5.5.1.1 Character Encodings for Czech

- iso_8859_2
- cp_1250
- utf_8, utf_16, ucs_4

5.5.1.2 Word Segmentation in Czech

The Czech segmenter follows all of the general segmentation rules in the white space languages.

5.5.1.3 Stemming in Czech

This section describes the standard stemmer and the expanded stemmer used for stemming in Czech.

5.5.1.3.1 Standard Stemmer

The standard Czech stemmer follows the general stemming rules as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Examples
Noun	tělem -> tělo, města -> město, výzkumy -> výzkum

Category	Examples
Verb	máš -> mít, vrátil -> vrátit, dostane -> dostat, pracuji -> pracovat
Adjective	velká -> velký, starší -> starý
Adverb	brzy -> brzy, dnes -> dnes

5.5.1.3.2 Expanded Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text, such as email, online documents, or queries. In Czech, this includes accented characters missing their diacritics and proper names without word-initial capitalization. For instance:

Case Variants

The expanded version accepts lower case letters in addition to capital letters for those words where the capitals are obligatory.

Example	Output
Slovensko	Slovensko
slovensko	Slovensko

Deaccented Characters

The expanded version also allows deaccented characters in place of accented ones.

Example	Output
padák	padák
padak	padák

5.5.1.4 Part-of-Speech Tagging in Czech

The following table shows the Czech tag set available for using in defining custom entities. The tag names are accompanied by a brief description and one or more examples. The tag set makes no distinction for gender.

Umbrella Tag	Complete Tag	Description	Examples
Adj	Adj	Adjectives	úřadní, úřadními, úřadním
	Adj-Comp	Comparative adjectives	úřadnější, úřadnějšími, úřadnějším
	Adj-Sup	Superlative adjectives	nejuřednější, nejuřednějším, nejuřednějšími
Adv	Adv	Adverbs	úředně,zúředněně, zdeúředně
	Adv-Comp	Comparative adverbs	úředněji, zúředněněji, zdeúředněji
	Adv-Sup	Superlative adverbs	nejúředněji, nejzúředněněji, nejzdeúředněji
Conj	Conj	Conjunction	či, čili, že
Interj	Interj	Interjection	úhuhu, ó, ólala
Nn	Nn	Invariant noun	sec, pH, um.
	Nn-Pl-Gen	Plural, genitive noun	úřadníků, úřeků, úřezů
	Nn-Pl-Case	Plural, nominative, vocative, accusative, dative, locative and instrumental noun	úřadníci, úřadníkům, úřadníků
	Nn-Sg-Gen	Singular, genitive noun	úřadníka, úřeku, úřezu
	Nn-Sg-Case	Singular, nominative, vocative, accusative, dative, locative and instrumental noun	úřadníka, úřadníkovi, úřadníkem
	Nn-Net	URI, email address	www.inxight.com, info@inxight.com,

Umbrella Tag	Complete Tag	Description	Examples
Prop	Prop	Proper noun	ČSLA, Ňasko, Ňasku, Ňaska, Ňaskem, Íliada
Num	Num	Number expression other than cardinal or ordinal	XV, mil.
	Num-Card	Cardinal number	dvě, dvěma, dvou
	Num-Ord	Ordinal number	šestýma, šestými, šestým
Part	Part	Particle	řekněmež, čau, žbluňk
Prep	Prep	Preposition	zmísta, zkraje, zaň
Pron	Pron-Dem-Pl	Plural demonstrative pronoun	týmiž, týmaž, týchž
	Pron-Dem-Sg	Singular demonstrative pronoun	týž, týmž
	Pron-Pl	Plural pronoun	číchkoliv, číchkoli, čímisi, čímasi, čímsi
	Pron-Sg	Singular pronoun	číhosi, čímsi, čímukoliv, čímukoli, číhokoliv, číhokoli, číhosi
	Pron-Int/Rel	Interrogative/relative pronoun	čí, čími, čím, čích, čímu
	Pron-Refl	Reflexive pronoun	svůj, svých, svýmu, svým
	Pron-Pers-Sg	Singular personal pronoun	on, ono, ona, ty, von
	Pron-Pers-Pl	Plural personal pronoun	vy, vás, vám, vámi
	Pron-Poss	Possesive pronoun	tvůj, váš, vaší

Umbrella Tag	Complete Tag	Description	Examples
V	V-Inf	Infinitive verb	dělat, úžit, úžívat
	V-Imp	Imperative verb	dělej, dělejme, dělejte
	V-Ind	Indicative, verb	dělána, dělány, dělánu, dělánu, dělána
	V-PaPart	Past participle	dělal, dělals, dělaly, dělali
	V-Inf-Be	Verb "to be", infinitive	být, bývat, nebýt, nebý-vat
	V-Imp-Be	Verb "to be", imperative mood	buď, budiž, buďme, buďte
	V-Pres-Be	Verb "to be", present tense	je, jest, jsi, jste, jsme, jsou, jsem
	V-Fut-Be	Verb "to be", future tense	bude, budu, budeš, budete, budou, budem
	V-PaPart-Be	Verb "to be", past par-ticiple	byl, byla, bylo, byla, byly, byli
	V-APart	Adjectival/adverbial participle	dělaje, dělajíc, dělajíce
	V-Aux	Auxiliary verb	by, bys, byste, bych, bychom
Punct	Punct-Sent	Sentence ending punc-tuation	! ? .
	Punct-Comma	Comma	,
	Punct-Open	Opening punctuation	(
	Punct-Close	Closing punctuation	)
	Punct-Quote	Quote	" '
	Punct	Other punctuation	+ -

5.5.1.4.1 Unfound Words

Words not found in the tagger dictionary are passed to the Czech guesser to be assigned the most likely tag. The Czech guesser assigns tags to unfound words based on a set of rules about Czech

morphology. For example, a word ending in **-ova** is likely an adjective. Internet and e-mail addresses are assigned the tag `Nn-Net`.

Capitalization information is also important; for instance, capitalized words tend to be guessed as proper nouns.

5.5.1.5 Grouping in Czech

Czech noun groups can be simple or compound nouns with their modifiers.

Modifiers can be adjectives or ordinal numbers but not determiners or pronouns.

Modifiers can have adverbs as their own modifiers.

For example:

- ministrem Vrbou
- úst paní ministryně Štěpové

5.5.2 Extraction

This section describes the extraction-specific information for Czech.

5.5.2.1 Predefined Entity Types

This section describes the predefined entity types supported by the Czech language module and examples of each.

5.5.2.1.1 NOUN_GROUP

Czech noun groups can be simple or compound nouns with their modifiers. Modifiers can be adjectives or ordinal numbers but not determiners or pronouns. Modifiers can have adverbs as their own modifiers.

For example:

- ministrem Vrbou
- úst paní ministryně Štěpové

5.6 Danish Language Reference

This chapter describes the behavior of the Danish language module.

5.6.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Danish texts, including word segmentation stemming, and tagging.

5.6.1.1 Character Encodings for Danish

- iso_8859_1
- cp_1252
- utf_8, utf_16, ucs_4

5.6.1.2 Word Segmentation in Danish

The Danish segmenter follows all of the general rules for word segmentation for the white space languages. The Danish segmenter has the following language-specific behavior.

The Danish segmenter keeps together plurals and possessives spelled with **s** or **'s**. Hyphens are not separated from compound parts written with a hyphen. Periods are not separated from ordinal digit expressions. Examples are shown below:

Text	Segmented
Eriks	Eriks
14.	14.

Text	Segmented
post- og telegrafvæsenet	post-
	og
	telegrafvæsenet

Related Topics

- Word Segmentation
- White Space Languages

5.6.1.3 Stemming in Danish

This section describes the standard stemmer and the expanded stemmer used for stemming in Danish.

5.6.1.3.1 Standard Stemmer

The Danish stemmer follows the general stemming rules as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. The base forms for Danish shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Baseform	Example
Noun	Indefinite singular	kager -> kage, cyklen -> cykel
Verb	Infinitive	sendes -> sende, luk -> lukke
Adjective	Base form	kolde -> kold, smukkest -> smuk
Adverb	Base form or source form	oftest ->ofte, bagfra -> bagfra

5.6.1.3.2 Expanded Inflectional Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text (such as email, online documents, or queries). See Expanded Inflectional Stemming for general behavior. The specifics for Danish follow.

Case Variants

The expanded version accepts lower case letters in addition to capital letters for those words where the capitals are obligatory.

Example	Output
erik	Erik
Erik	Erik

Typewriter Forms of Accented Letters

The expanded version accepts typewriter conventions for accented letters. That is, **å** is recognized when written as **aa**, **æ** when written as **ae**, and **ø** when written as **oe**.

Example	Output
blaa	blå
blå	blå

Deaccented Characters

The expanded version also allows deaccented characters in place of accented ones.

Example	Output
bla	blå
blå	blå

Hyphenation

Hyphens in non-numeric expressions are optional in the expanded version. This is useful for text that has been pre-processed to undo line-breaking hyphenation by deleting both the line break and the hyphen, without regard to whether the hyphen is a genuine part of the word or only there for the line break.

Example	Output
Vdag	V-dag
V-dag	V-dag

5.6.1.4 Part-of-Speech Tagging in Danish

The following table shows the Danish tag set available for using in defining custom entities. The tag names are accompanied by a brief description and one or more examples. If the example consists of more than one word, the word exemplifying the current tag is in **bold**. The tag set makes no distinction for gender.

Umbrella Tag	Complete Tag	Description	Examples
`Abbr`	`Abbr`	Abbreviation	nov., kg, dkr., USA
`Adj`	`Adj`	Adjective	det **gode** brød
	`Adj-Comp`	Comparative adjective	et **bedre** forslag
	`Adj-Gen`	Genitive adjective	den **enkeltes** tryghed
	`Adj-PaPart`	Past participle used as adjective	**bestemt** niveau
	`Adj-PaPart-Gen`	Past participle used as adjective, genitive	den **ansattes** papirer
	`Adj-PrPart`	Present participle adjective	manglende

Umbrella Tag	Complete Tag	Description	Examples
Adv	Adv	Adverb (includes particles)	igen
	Adv-Comp	Comparative adverb	tidligere
	Adv-Int/Rel	Interrogative or relative adverb	hvor, hvorefter
	Adv-Sup	Superlative adverb	oftest

Umbrella Tag	Complete Tag	Description	Examples
Aux/V	Aux/V-Infin-blive	Infinitive auxiliary or main verb **blive**	blive
	Aux/V-Infin-faa	Infinitive auxiliary or main verb **få**	få
	Aux/V-Infin-have	Infinitive auxiliary or main verb **have**	have
	Aux/V-Infin-vaere	Infinitive auxiliary or main verb **vaere**	være
	Aux/V-PaPart-blive	Past participle auxiliary or main verb **blive**	blevet
	Aux/V-PaPart-faa	Past participle auxiliary or main verb **få**	fået
	Aux/V-PaPart-have	Past participle auxiliary or main verb **have**	haft
	Aux/V-PaPart-vaere	Past participle auxiliary or main verb **vaere**	været
	Aux/V-Past-blive	Past tense auxiliary or main verb **blive**	blev
	Aux/V-Past-faa	Past tense auxiliary or main verb **få**	fik
	Aux/V-Past-have	Past tense auxiliary or main verb **have**	havde
	Aux/V-Past-vaere	Past tense auxiliary or main verb **vaere**	var
	Aux/V-Pres-blive	Present tense auxiliary or main verb **blive**	bliver
	Aux/V-Pres-faa	Present tense auxiliary or main verb **få**	får
	Aux/V-Pres-have	Present tense auxiliary or main verb **have**	har
	Aux/V-Pres-vaere	Present tense auxiliary or main verb **vaere**	er
Cmpd	Cmpd-Part	Left compound part	**post** - og tele-grafvæsenet

Umbrella Tag	Complete Tag	Description	Examples
Conj	Conj	Conjunction	at, når
	Conj-Coord	Coordinating conjunction	og, eller
	Conj-hvis	Conjunction or relative pronoun **hvis**	hvis
	Conj-som	Conjunction or relative pronoun **som**	som
Det	Det	Determiner	en
	Det-Indet	Indeterminate determiner	forskellig, somme
	Det-Indet-Gen	Indeterminate determiner, genitive	forskelliges
	Det-Coord	Conjunctional adverb	både
	Det/Pron-Int/Rel	Interrogative or relative pronoun	hvad, hvem, hvilke
	Det/Pron-Poss	Possessive determiner or pronoun	vores, min
	Det/Pron-Poss-Refl	Reflexive possessive pronoun	sin, sit, sine
	Det/Pron-Quant	Quantifying determiner or pronoun	mange
	Det/Pron-Quant-Comp-mere	Comparative **mere**	mere
	Det/Pron-Quant-Gen	Genitive quantifying determiner or pronoun	manges
	Det/Pron-Quant-Pre	Quantifying pre-determiner or pronoun	alle, hver
	Det/Pron-Quant-Sup-mest	Superlative **mest**	mest
Func	Func	Function word (miscellaneous category)	ambulatorie, barne
Interj	Interj	Noun	kvinde

Umbrella Tag	Complete Tag	Description	Examples
Nn	Nn	Genitive noun	kvindens
	Nn-Gen	Lowercase and uppercase letters	b, N
	Nn-Letter	URL and e-mail address	www.inxight.com info@inxight.com
	Nn-Net		
Num	Num	Cardinal number (in digits or words)	3m tre
Ord	Ord	Ordinal number, in digits or spelled out	20., femte
Part	Part-Inf	Infinitival particle **at**	få lov **at** indtage
	Part-Neg	Negative particle	ikke
Prep	Prep	Preposition	med, hos
	Prep-af	Preposition **af**	af
Pron	Pron	Pronoun	den, denne
	Pron-Expl	Expletive pronoun	**der** var 400 deltagere
	Pron-Gen	Genitive pronoun	begges
	Pron-Pers	Personal pronoun	jeg, mig
	Pron-Recip	Reciprocal pronoun	hinanden
	Pron-Recip-Gen	Genitive reciprocal pronoun	hinandens
	Pron-Rel	Relative pronouns **der** and **som**	familier, **der** skilles
Prop	Prop	Proper name, initials or title	Ole, H., fru, dr.
	Prop-Gen	Genitive proper name	**Jensens** bil

Umbrella Tag	Complete Tag	Description	Examples
Punct	Punct	Miscellaneous punctuation	-)
	Punct-Comma	Comma	,
	Punct-Sent	Sentence boundary punctuation	. ? !
V	V-Impv	Imperative verb	skriv
	V-Infin	Infinitive verb	skrive
	V-PaPart	Past particple verb	skrevet
	V-Past	Past tense verb	skrev
	V-Past-SForm	Past tense S-form verb	taltes
	V-Pres	Present tense verb	sker
	V-Pres-SForm	Present tense S-form verb	sendes

5.6.1.5 Grouping in Danish

A Danish simple noun phrase is a noun or series of nouns, optionally modified by a possessive form of a proper name, adjectives, or ordinal numbers. Possessive pronouns are not included in noun groups. For example:

- varmt vand
- fin, ny cykel
- Odenses vedkommende
- 29. october

Adjectives and nouns may be joined by coordinating conjunctions like **og** 'and' and **eller** 'or'. Series of compound parts joined to a noun by a coordinating conjunction are also allowed.

- vand og salt
- stor eller lille is
- vand- og varmemester

A simple noun phrase may be followed by a prepositional phrase beginning with **af**. For instance:

- ejeren af hesten

5.6.2 Extraction

This section describes the extraction-specific information for Danish.

5.6.2.1 Predefined Entity Types

This section describes the predefined entity types supported by the Danish language module and examples of each.

5.6.2.1.1 NOUN_GROUP

A Danish simple noun phrase is a noun or series of nouns, optionally modified by a possessive form of a proper name, adjectives, or ordinal numbers. Possessive pronouns are not included in noun groups. For example:

- varmt vand
- fin, ny cykel
- Odenses vedkommende
- 29. october

Adjectives and nouns may be joined by coordinating conjunctions like **og** 'and' and **eller** 'or'. Series of compound parts joined to a noun by a coordinating conjunction are also allowed.

- vand og salt
- stor eller lille is
- vand- og varmemester

5.7 Dutch Language Reference

This chapter describes the behavior of the Dutch language module.

5.7.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Dutch texts, including word segmentation and stemming.

5.7.1.1 Character Encodings for Dutch

- iso_8859_1
- cp_1252
- utf_8, utf_16, ucs_4

5.7.1.2 Word Segmentation in Dutch

The Dutch segmenter follows all of the general segmentation rules in the white space languages.

The Dutch segmenter has the following language-specific behavior. It does not split contractions. Plurals and possessives spelled with s or 's are not split. Hyphens are not separated from compound parts written with a hyphen.

Text	Segmented
m'n	m'n
'k	'k
auto's	auto's
Jansens	Jansens

Text	Segmented
honden- en kattenvoer	honden-
	en
	kattenvoer

Related Topics

- Word Segmentation
- White Space Languages

5.7.1.3 Stemming in Dutch

This section describes the standard stemmer and the expanded stemmer used for stemming in Dutch.

5.7.1.3.1 Standard Stemmer

The Dutch stemmer follows the general stemming rules as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Baseform	Examples
Noun	Non-diminutive singular	bloem -> bloem, emmers -> emmer, kinderen -> kind
Verb	Infinitive	schrijft -> schrijven, hebt -> hebben
Adjective	Base form	lange -> lang, onhandigste -> handig
Adverb	Base form or source form	eventjes -> even, liefst -> graag, gisteren -> gisteren

5.7.1.3.2 Expanded Inflectional Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text (such as email, online documents, or queries). See Expanded Inflectional Stemming for general behavior. Here, we list the specifics for Dutch.

The expanded version of the Dutch stemmer does not require correct capitalization and accentuation and allows required hyphens to be optional. It also accepts certain nonstandard conventions for hyphens in compounds.

Hyphenation

Hyphens are officially used in compounds only when the first compound element ends in a vowel and the second element starts with a vowel, in order to facilitate the pronunciation. The expanded version accepts these compounds even when obligatory hyphens are missing.

Example	Output
auto-ongeluk	auto | ongeluk
autoongeluk	auto | ongeluk

Most compounds are usually written without hyphens, but in common practice many compounds are often written both with and without hyphens. The expanded version allows optional hyphenation in non-vowel environments.

Example	Output
kinderbioscoop	kind | bioscoop
kinder-bioscoop	kind | bioscoop

Case Variants

The expanded version accepts lower case letters in addition to capital letters for those words where the capitals are obligatory.

Example	Output
USA	USA

Example	Output
usa	USA

Deaccented Characters

The expanded version allows deaccented characters in place of accented ones.

Example	Output
privé	privé
prive	privé

5.7.1.3.3 Dutch Compound Analysis

Compounding in Dutch can combine various parts of speech: nouns can combine with nouns, nouns with adjectives, and verbs with nouns.

Note:
The sample output below uses the vertical bar (|) to delimite terms or stems. Compounds are always broken up.

Noun-Noun

Dutch noun compounds often incorporate linking elements. For instance, when the words **leven** and **echt** combine to form a compound, the linking element -**s-** is added between them, giving **levensecht**. The two most frequent linking elements are **-s**- and **-en-**. (A third linking element, -**e-**, occurs in only a few irregular compounds.)

Example	Output
begrafenisstoet	begrafenis | stoet
mensenrechtenorganisaties	mens | recht | organisatie
levensecht	leven | echt

Nouns with vowel changes or other irregularities are also handled. For example:

Example	Output
scheepskapitein	schip | kapitein
zonnescherm	zon | scherm

If the first compound element ends in a vowel and the second element starts with a vowel, a hyphen is conventionally inserted between the elements in the source form. This hyphen does not appear in the stemmer output.

Example	Output
auto-ongeluk	auto | ongeluk

Compounds like **boeken- en platenzaak** ("book and record shop") or **kindertheater en -bioscoop** ("children's theater and cinema") are sometimes seen. The hyphen in the first part **boeken-**replaces the noun **zaak**, and that in**-bioscoop** replaces **kinder**.

Example	Output
boeken- en platenzaak	boek en platenzaak
kindertheater en -bioscoop	kind | theater en bioscoop

Verb-Noun

Compounds can also combine verbs and nouns. The verb part is stemmed to the infinitive form of the verb.

Example	Output
schrijfwijze	schrijven | wijze

Noun-Adjective

In compounds combining adjectives and nouns, the linking elements often seen in noun-noun compounds are absent.

Example	Output
kinderloos	kind | loos

5.7.1.4 Part-of-Speech Tagging in Dutch

The following table shows the Dutch tag set available for using in defining custom entities. The tag names are accompanied by a brief description and one or more examples. If the example consists of more than one word, the word exemplifying the current tag is in **bold**. The tag set makes no distinction for gender.

Umbrella Tag	Complete Tag	Description	Examples
`Adj`	`Adj-Attr`	Attributive adjective	een **snelle** auto
	`Adj-Post`	Postmodifying adjective	wat **anders**
	`Adj-Pred`	Predicative adjective	hij rijdt **snel**
`Adv`	`Adv`	Non-adjectival adverb	stroomopwaarts
	`Adv-Deg`	Adverbs that can modi-fy adjectives	hij wil **te** snel
	`Adv-Int`	Interrogative adverb	**waarom** gaat hij
	`Adv-Pron`	Pronominal adverb	hij praat **hierover**
`Aux`	`Aux-Fin`	Finite auxiliary verb	hij **is** geweest
	`Aux-Inf`	Infinitive auxiliary verb	hij zal **zijn**
	`Aux-PaPart`	Past participle auxiliary verb	hij is **geweest**

Umbrella Tag	Complete Tag	Description	Examples
Cmpd	Cmpd-Left	Left truncated part of compound	**honden** - en kattenvoer
	Cmpd-Right	Right truncated part of compound	kattenvoer en - **melk**
Conj	Conj-Comp	Comparative conjunction	zo groot **als**
	Conj-Coord	Coordinating conjunction	jan **en** marie
	Conj-Inf	Infinitive conjunction	door **te** vragen
	Conj-Rel	Relative conjunction	het kind **dat** ...
	Conj-Sub	Subordinating conjunction	**hoewel** hij er was
	Conj-Sub-Adv	Interrogative adverb or subordinate conjunction	**wanneer** gaat hij weg?
Det	Det-Art	Determiner	**een** bus
	Det-Dem	Demonstrative determiner	**deze** machine gaat goed
	Det-Indef	Indefinite determiner	**geen** broer
	Det-Int/Rel	Interrogative or relative determiner	de vraag **wier** man ...
	Det-Poss	Possessive determiner	**mijn** boek
	Det-Post-Indef	Indefinite postdeterminer	de **beide** broers
	Det-Pre-Indef	Indefinite predeterminer	binnen **al** deze pakketten
Interj	Interj	Interjections	och
Nn	Nn	Common noun	boek
	Nn-Letter	Lowercase and uppercase letters	b, N
	Nn-Net	URL and e-mail address	www.inxight.com info@inxight.com

Umbrella Tag	Complete Tag	Description	Examples
Num	Num	Cardinal number	125, vijf, 12/2
Ord	Ord	Ordinal number	vijfde, 125ste, 12de
Part	Part-Inf	Particle of Dutch **'te+infinitive'** construction	hij hoopt **te** gaan
	Part-Neg	Negation particle	hij gaat **niet** snel
	Part-Prefix	Separated prefix of (pronominal) adverb or verb	hij loopt **mee**
Prep	Prep	Preposition	in
	Prep-Circ	Right part of circumposition	tot nu **toe**
	Prep-Post	Postposition	veel passanten **langs** komen
	Prep-van	Preposition **van**	van
Pron	Pron-Dem	Demonstrative pronoun	**deze** gaat goed
	Pron-Indef	Indefinite pronoun	beide
	Pron-Int/Rel	Interrogative or relative pronoun	de vraag **wie** ...
	Pron-Pers	Personal pronoun	hij
	Pron-Rel	Relative pronoun	de man **die** lachte
Prop	Prop	Proper noun, including initials and title of address	Peter, C., Prof.
	Prop-Art	Article beginning a name	**De** Vries
	Prop-Prep	Preposition beginning a name	**Van** den Broek

Umbrella Tag	Complete Tag	Description	Examples
Punct	Punct	Miscellaneous punctuation	{ } [ ] - ---
	Punct-Comma	Comma	,
	Punct-Quote	Quotation type punctuation (includes parentheses)	" ' ' ( )
	Punct-Sent	Sentence final punctuation	. ? ! ;
	Punct-Slash	Slash mark	/
V	V-Fin	Finite verb	zegt
	V-Inf	Infinitive verb	zeggen
	V-PaPart	Past participle verb	gezegd
	V-PrPart	Present participle verb	zeggend

5.7.1.5 Grouping in Dutch

Dutch noun phrases consist of a noun with optional modifiers, such as adjectives, as in:

- Amerikaanse minister

Compounds are also grouped in Dutch. The compound parts may be modified, and there may be more than one, separated by commas and/or conjunctions.

- boeken- en platenzaak
- kindertheater en -bioscoop

In Dutch, noun coordination is allowed, for instance with **en** 'and', as in:

- productiviteitscijfers en fabrieksbestellingen
- specifieke juwelen en kledingstukken
- studenten, ouders en leraren

The only preposition included in noun groups is van, as shown below, although **ter** is also allowed in names. Names may begin with **Van**, **De**, **Den**, **Der**, or **Ter**; this is the only time a preposition or determiner is allowed to occur at the beginning of a noun phrase.

- voorstel van de werkgevers
- militaire nederlaag van de afgelopen weken

5.7.2 Extraction

This section describes the extraction-specific information for Dutch.

5.7.2.1 Predefined Entity Types

This section describes the predefined entity types supported by the Dutch language module and examples of each.

5.7.2.1.1 NOUN_GROUP

Dutch noun phrases consist of a noun with optional modifiers, such as adjectives, as in:

- Amerikaanse minister

Compounds are also grouped in Dutch. The compound parts may be modified, and there may be more than one, separated by commas and/or conjunctions.

- boeken- en platenzaak
- kindertheater en -bioscoop

In Dutch, noun coordination is allowed, for instance with **en** 'and', as in:

- productiviteitscijfers en fabrieksbestellingen
- specifieke juwelen en kledingstukken
- studenten, ouders en leraars

5.8 English Language Reference

This chapter describes the behavior of the English language module.

5.8.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of English texts, including word segmentation, stemming, and tagging.

5.8.1.1 Character Encodings for English

- iso_8859_1
- cp_1252
- utf_8, utf_16, ucs_4

5.8.1.2 Word Segmentation in English

The English segmenter follows all of the general segmentation rules in the white space languages. The English segmenter has the following language-specific behavior.

In English, contractions like **don't**, **can't** and **won't** are separated into their constituent syntactic units. **Ain't** is not separated, since there is no clearly correct way to break it. The possessive endings **'s** and **'** are separated from the words they modify.

Text	Segmented
can't	can
	n't
won't	will
	n't

Text	Segmented
it's	it
	's
ain't	ain't
helper's	helper
	's
helpers'	helpers
	'

Abbreviations are not split from their punctuation, but do get split from following hyphens. Hyphens that occur in between two abbreviations will not break the syntactic unit. Abbreviations are listed in a system dictionary as well as in a set of rules allowing for uppercase and lowercase letters as well as periods and optional hyphens.

Combinations of alphabetic, numeric, and optionally, punctuation characters are kept together. For example:

Text	Segmented
Apr.-	Apr.
	-
D-Nebr.	D-Nebr.
3a.m.	3a.m.
11Jan.	11Jan.
Mon.-Thurs.	Mon.-Thurs.

Text	Segmented
Bloomberg-U.S.	Bloomberg-U.S.

Related Topics

- Word Segmentation
- White Space Languages

5.8.1.3 Stemming in English

This section describes the standard stemmer and the expanded stemmer used for stemming in English.

5.8.1.3.1 Standard Stemmer

The English stemmer follows the general stemming rules as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below.

Category	Baseform	Examples
Noun	Singular	dog, dogs -> dog
Verb	Infinitive	runs, ran, run -> run
Adjective	Base form	happy, happier, happiest -> happy
Adverb	Base form or source form	quickly -> quickly

English pronouns are stemmed in the following way. All uninflecting forms stem to themselves. Plural-only forms and all personal pronouns maintain their number and gender information. If applicable, these pronouns are stemmed to the nominative form. All other forms stem to the singular form. This is shown in the table below:

Text	Stem
none	none
that	that
themselves	themselves
her	she
these	this

The standard stemmer handles the spelling variation found in American and British English. Both variants stem to the American spelling. These behaviors are shown in the following table:

Text	Stem
color	color
colour	color
organization	organization
organisation	organization

5.8.1.3.2 Expanded Inflectional Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text (such as email, online documents, or queries). See Expanded Inflectional Stemming for general behavior. Following is a list of the specifics for English.

Case Variants

The expanded version accepts lower case letters in addition to capital letters for words that are usually capitalized. If both lower and upper case variants are included in the stemmer, both are returned as stems. A lower case variant returns an uppercase stem if this is the only one included in the stemmer.

Example	Output
Eric	Eric
eric	Eric

Hyphenation

To aid software that handles line-breaking hyphens by deleting them and concatenating the two parts of the broken word, hyphens in non-numeric expressions are optional in the expanded version, so that words that are truly hyphenated will still be recognized.

Example	Output
square-dance	square-dance
squaredance	square-dance
motherinlaw	mother-in-law

5.8.1.3.3 Derivational Stemmer

The derivational stemmer is designed to produce the root word for an entry, crossing word categories when necessary. For example, the noun **connection** is derived from the verb **connect** by adding the suffix **-ion**.

Therefore, the derivational stemmer finds the root **connect** for the noun **connection**. Similarly, **driver** is stemmed to **drive** and **quickly** to **quick**.

Text	Stem
connection	connect
belongings	belong
driver	drive
quickly	quick

5.8.1.3.4 Inflectional Stemmer Guesser

The inflectional stemmer guesser contains a set of morpological rules that can apply to words that are unknown to the standard or expanded inflectional stemmer and therefore cannot be stemmed.

Linguistics processing first attempts to perform stemming using the standard or expanded inflectional stemmer, and then applies the stemmer guesser only to words that cannot be conventionally stemmed.

5.8.1.4 Part-of-Speech Tagging in English

The following table shows the English tag set. The tag names are accompanied by a brief description and one or more examples. If the example consists of more than one word, the word exemplifying the current tag is in **bold**.

Umbrella Tag	Complete Tag	Description	Examples
`Abbr`	`Abbr`	Abbreviation	i.e.
`Adj`	`Adj`	Adjective	big
	`Adj-Comp`	Comparative adjective	bigger
	`Adj-Ord`	Ordinal adjective	third
	`Adj-Sup`	Superlative adjective	biggest
`Adv`	`Adv`	Adverb	quickly
	`Adv-Comp`	Comparative adverb	sooner
	`Adv-Int/Rel`	**wh-** adverb	how
	`Adv-Sup`	Superlative adverb	soonest
`Aux`	`Aux`	Auxiliary or modal	could
`Conj`	`Conj-Coord`	Coordinating conjunction	and
	`Conj-Sub`	Subordinating conjunction	unless

Umbrella Tag	Complete Tag	Description	Examples
Det	Det	Invariant determiner (singular or plural)	**some** food
	Det-Def	Definite determiner	the
	Det-Indef	Indefinite determiner	an
	Det-Int	Interrogative determiner	**what** time?
	Det-Int/Rel	Interrogative or relative determiner	whose
	Det-Pl	Plural determiner	**those** apples
	Det-Poss	Possessive determiner	my
	Det-Rel	Relative determiner	whatsoever
	Det-Sg	Singular determiner	every
Interj	Interj	Interjection	oh, hello
Nn	Nn	Invariant noun	sheep
	Nn-Letter	Letter	b, N
	Nn-Net	URL, e-mail address	www.inxight.com, info@inxight.com
	Nn-Pl	Plural noun	computers
	Nn-Sg	Cardinal number or other numeric expression	farmer
Num	Num	Cardinal number or other numeric expression	40.5, 11/27/00, $12.55, 12%, xvii, 9:00
Part	Part-Inf	Infinitive marker	**to** be or not to **be**
	Part-Neg	Negative particle	not
	Part-Poss	Possessive marker	John**'s** coat

Umbrella Tag	Complete Tag	Description	Examples
Prep	Prep	Preposition	below
	Prep-at	Preposition **at**	at
	Prep-of	Preposition **of**	of
Pron	Pron	Pronoun	he
	Pron-Int	**wh** pronoun	**what** do you want?
	Pron-Int/Rel	**wh** pronoun	who
	Pron-Refl	Reflexive pronoun	himself
	Pron-Rel	Relative pronoun	whoever
Prop	Prop	Name of a person or thing	Graceland
Punct	Punct	Other punctuation	- ; /%$
	Punct-Close	Closing punctuation	)] }
	Punct-Comma	Comma	,
	Punct-Open	Opening punctuation	([{
	Punct-Quote	Quote	" "
	Punct-Sent	Sentence-ending punctuation	. ! ?

Umbrella Tag	Complete Tag	Description	Examples
V	V-Inf-be	Infinitive **to be**	be
	V-PaPart	Verb, past participle,**-ed** verb form	has **walked**
	V-PaPart-be	Past participle of **to be**	has **been**
	V-PaPart-have	Past participle of **to have**	he has **had**
	V-Past	Verb, past tense	ran
	V-Past-have	Past tense of have	we **had**
	V-Past-Pl-be	Verb, past tense plural of **to be**	were
	V-Past-Sg-be	Verb, past tense singular of **to be**	was
	V-Pres	Verb, present tense or infinitive	sit
	V-Pres-3-Sg	Verb, present tense, 3rd person singular	sits
	V-Pres-3-Sg-have	Present tense, 3rd person singular of have	has
	V-Pres-have	Present tense or infinitive of **have**	have
	V-Pres-Pl-be	Verb, present tense plural of **to be**	are
	V-Pres-Sg-be	Verb, present tense singular of **to be**	is
	V-PrPart	Verb, present participle, **-ing** verb form	is **walking**

5.8.1.4.1 Unfound Words

Words not found in the tagger dictionary are passed to the English tagger guesser to be assigned the most likely tag. The English tagger guesser assigns tags to unfound words based on a set of rules about English morphology, for example, a word ending in **-ly** is likely an adverb. Internet and e-mail addresses are assigned the tag `Nn-Net`.

Capitalization information is also important; for instance, capitalized words tend to be guessed as proper nouns. Combinations of alphabetic and numeric characters are guessed as proper nouns as well. Ordinal numbers are tagged either as noun or adjective, depending on the context as determined by the software.

5.8.2 Extraction

This section describes the extraction-specific information for English.

5.8.2.1 Advanced Parsing

The extraction process performs linguistic processing by using tools that include semantic and syntactic knowledge of words. In general, linguistic processing identifies paragraphs, sentences, and clauses, and then identifies semantic and syntactic information within the text. Extraction provides two modes for linguistic processing in English: standard and advanced. The default is standard.

Advanced parsing offers richer, better coordinated noun phrase extraction that includes syntactic function attributes, as well as pronominal resolution and is available when processing extraction rules only.

5.8.2.2 English Subtypes

English supports subtypes in the types `FACILITY`, `GEO_AREA`, `GEO_FEATURE`, `NIN`, `ORGANIZATION`, `REGION`, `SOCIAL_MEDIA`, `URI`.

Related Topics

- Subtypes

5.8.2.3 Predefined Entity Types

This section describes the predefined entity types supported by the English language module and examples of each.

Click any link to jump to that subsection: ADDRESS1 and ADDRESS2, CONTINENT, COUNTRY, CURRENCY, DATE, DAY, FACILITY, GEO_AREA, GEO_FEATURE, GEOCOORD, HOLIDAY, LANGUAGE, LOCALITY, MEASURE, MGRS, MONTH, NAME_DESIGNATOR, NIN, NOUN_GROUP, ORGANIZATION, PEOPLE, PERCENT, PERSON , PHONE, PRECURSOR, PRODUCT, PROP_MISC, REGION, SOCIAL_MEDIA, TICKER, TIME, TIME_PERIOD, TITLE, URI, VEHICLE, WEAPON, and YEAR.

Note:
The common mentions extraction pack enables extraction of a number of common nouns in English. See Common Mentions Content.

5.8.2.3.1 ADDRESS1 and ADDRESS2

The format for `ADDRESS1` and `ADDRESS2` is based on US address forms.

`ADDRESS1` corresponds to the first part of an address. At least two pieces of the following address information must be adjacent to be extracted as `ADDRESS1`:

- Building name
- House number
- Street name
- Apartment number
- PO Box and number
- Neighborhood

`ADDRESS2` is meant to annotate the second part of an address. At least two pieces of the following address information must be adjacent to be extracted as `ADDRESS2`:

- Postal code
- City
- State/province/department

Country names are always extracted as `COUNTRY`, even in an address context.

Text	ADDRESS1	ADDRESS2
1234 Mahana St. Honolulu Hawai‘i 96816	1234 Mahana St.	Honolulu Hawai‘i 96816
PO Box 10101	PO Box 10101	
Rural Route 5	Rural Route 5	

Text	ADDRESS1	ADDRESS2
Paris, Texas		Paris, Texas

5.8.2.3.2 CONTINENT

Any of the continents, for example:

- Asia
- Europe

Note:
America and Australia are extracted as `COUNTRY` only.

5.8.2.3.3 COUNTRY

Names of countries, and abbreviations. This list also includes the names of geo-political entities for which the conventional labels do not apply, such as disputed territories or territories that have not been internationally recognized:

- Italy
- U.K.
- USA
- Palestinian National Authority
- Taiwan

5.8.2.3.4 CURRENCY

Quantities of world currency, and ranges of amounts of currency:

- 35 cents
- 1.19 dlrs
- one dollar and twenty-five cents
- 785 to 995 dlrs

5.8.2.3.5 DATE

Dates are minimally composed of a number and month:

- April 2
- 26 November 1998

- September tenth
- fourth of June

5.8.2.3.6 DAY

Days of the week, including abbreviations:

- Monday
- Mon.
- TUES

5.8.2.3.7 FACILITY

Man-made structures, extracted as one of the following subtypes:

- `AIRPORT`–The names of primarily man-made or man-maintained structures whose primary use is as air transportation terminals. For example:
  - Los Angeles International Airport
  - South Capitol Street Heliport
- `BUILDGROUNDS`–The names of architectural and civil engineering structures, and outdoor spaces that are mainly man-made or man-maintained. There is no distinction with respect to their function, they could be civil or military facilities, they could be used for work or entertainment, or they could be monuments. For example:
  - Berlin Wall
  - Disneyland
  - Fort Knox
  - Grand Central Station
  - Statue of Liberty
- `PATH`–The names of primarily man-made or man-maintained structures that allows fluids, energy, persons, animals, or vehicles to pass from one location to another. For example:
  - Champs-Elysees
  - Erie Canal
  - London Bridge
  - Times Square
- `PLANT`–The names of facilities composed of one or more buildings used for industrial purposes. For example:
  - San Onofre Nuclear Generating Station
  - Three Mile Island

- `SUBAREA`–The names of portions of facilities, typically architectural ones, that are able to contain people, animals, or objects. For Example:
  - Air Canada Maple Leaf Lounge

5.8.2.3.8 GEO_AREA

A geographical area that captures a significant land mass, such as a group of countries, extracted as one of the following subtypes:

- `DOMESTIC`–The names of locations that do not cross national borders. For example:
  - Northern Illinois
  - South Florida
  - Midwest
- `INTL`–The names of locations that cross national borders. For example:
  - Southeast Asia
  - Western Europe

5.8.2.3.9 GEO_FEATURE

A non-artificial geographical location, that does not constitute a political entity extracted as one of the following subtypes:

- `BOUNDARY`–The names of locations such as borders. For example:
  - Mason-Dixon
  - Tropic of Cancer
- `CELESTIAL`–The names of astronomical locations that are outside of the boundaries of the Earth. For example:
  - Neptune
  - Mars
- `LAND`–The names of locations that are geologically or ecosystemically designed, non-artificial locations. For example:
  - Grand Canyon
  - Mount Fuji
- `WATER`–The names of locations that are bodies of water. For example:
  - Pacific Ocean
  - Lake Michigan
  - Volga River

5.8.2.3.10 GEOCOORD

Geographic coordinates, of various formats:

- 1234N/12345E
- LAT. 12.34N LONG. 012.34W
- 234500S/0123400W
- 12'34.5N4-012'34.5E6
- 3074N04429E
- 33 40' 56.14" N 69 56' 20.20" E
- 38˚53'23"N , 77˚00'27"W

5.8.2.3.11 HOLIDAY

Holidays when banks and businesses are closed (i.e., bank holidays) in the countries where English is the official language. Spelled out dates that are also names of holidays are extracted as HOLIDAY (e.g. "Fourth of July") whereas numeric dates that coincide with a holiday are extracted as DATE (e.g., "4th of July"):

- New Year's Day
- Fourth of July
- Martin Luther King Day

5.8.2.3.12 LANGUAGE

The name of a language.

- He speaks **French.**
- A book written in **Arabic**

5.8.2.3.13 LOCALITY

Name of a city, including abbreviations for major cities:

- Cairo
- New Delhi
- Honolulu
- N.Y.C.
- Seville

5.8.2.3.14 MEASURE

Any measurement, such as weight, volume, or length, in English or metric units, including standard abbreviations of measurement units:

- 25 cubic feet
- 20 grams
- 6m

Rates of change, and ratios and ranges of measurements:

- 65 mph
- 33 mpg
- five cts per share
- 20 dlrs per unit

5.8.2.3.15 MGRS

Military Grid Reference System coordinates, of various formats:

- 18SUH6743
- 42S VB 7917 2559

5.8.2.3.16 MONTH

Months of the year, including abbreviations:

- January
- Feb.
- OCT

5.8.2.3.17 NAME_DESIGNATOR

A designator that appears before a person's name:

- Attn: in "Attn: John Smith"
- c/o in "c/o John Smith"
- CC: in "CC: Rob Brown"

5.8.2.3.18 NIN

National identification number, extracted as the following subtype:

- US_SSN – the United States Social Security Number:
  - 012-44-5668

Note:
Any extracted `NIN/US_SSN` entities can be parsed and standardized using the Data Cleanse transform of Data Services by mapping them to one of the `SSN` input fields.

5.8.2.3.19 NOUN_GROUP

English noun groups are nouns with modifying adjectives. For example:

- biggest problem
- interest rate
- mortgage interest tax relief

5.8.2.3.20 ORGANIZATION

Commercial, governmental, educational, legal, and service agencies, including non-profit organizations, fine arts groups, and other associations and institutions, extracted as one of the following subtypes:

- `COMMERCIAL`–The name of commercial organizations, such as major companies or corporations. For example:
  - Apple Corporation
  - General Electric Co.

  Also, variants and abbreviations for companies or corporations:
  - Apple
  - IBM

  Note:
  Any extracted `ORGANIZATION/COMMERCIAL` entities can be parsed and standardized using the Data Quality Data Cleanse transform by mapping them to one of the `FIRM` input fields.
- `EDUCATIONAL`–The names of institutions focused primarily on education. For example:
  - Brown
  - Cambridge University
  - MIT
  - Stanford University
- `ENTERTAINMENT`–The name of any organization, whether nonprofit or for profit, whose main interest is in the production of performing arts material or events: dance, music, opera, theatre, magic, spoken word, circus arts, and musical theatre. For example:
  - Cirque du Soleil
  - Boston Symphony Orchestra
- `GOVERNMENT`–The names of organizations related to government, politics, or the state. For example:

  - Foreign Ministry
  - Air National Guard
- `MEDIA`–The names of any organization, whether nonprofit or for profit, at any level, whose main interest is in the production of publishing and/or radio-, TV-, cable-, satellite-, or web-broadcasting material. For example:
  - Associated Press
  - NBC
  - PBS
- `MEDICALSCIENCE`–The names of organizations focused on medical care or research. For example:
  - American Medical Association
  - Dana-Farber Cancer Institute
  - European Space Agency
- `RELIGIOUS`–The names of organizations focused on religion. For example:
  - Church of Jesus Christ of Latter Day Saints
  - Church of England
- `SPORTS`–The names of organizations focused on sports. For example:
  - Red Sox
  - New York Yankees
- `OTHER`–Any organization that does not fit into a more specific subtype. For example:
  - European Community
  - Benelux
  - Greenpeace
  - United Nations
  - EU federation

5.8.2.3.21 PEOPLE

Names referring to identifiable groups of people based on country, ethnicity, region, or religion. For example:

- Arabs
- Scots

5.8.2.3.22 PERCENT

A percentage:

- 220%
- 18 pc
- fifty percent

Percent expressions:

- from 10% to 20%
- between 5 and 10 percent

5.8.2.3.23 PERSON

An individual specified by name. A variety of forms are identified:

- Bill Clinton
- William J. Clinton
- William Jefferson Clinton
- Mustafa Al-Jaziri `Abd Al-Rahaman Nudle
- Mary Beth Josephine Thomas
- Ms. Washington
- Mr. Copperfield

Note:
Given and family names that occur by themselves are extracted as `PERSON` as long as they are not ambiguous with common names, with the exception of famous cases, such as Bush.

5.8.2.3.24 PHONE

Phone numbers based on US format:

- 1-408-738-6200
- 408-738-6200
- 738-6200
- (408) 738-6200
- 1-888-FLOWERS
- 408-738-6200 x111

International phone numbers based on French, German and Spanish formats:

- 11 11 22 22 22

- 11/22/33/44/55
- (01) 11 22 33 44 55
- (+49)-111-22-33333
- Telefon: 0111-22222

5.8.2.3.25 PRECURSOR

Weapon precursors, extracted as one of the following subtypes:

- `CHEMICAL`–Names of chemical substances that have been officially identified as used in the manufacturing of chemical weapons:
  - Cyanide
  - Dimethyl methylphosphonate
- `NUCLEAR`–Names of nuclear and radiological substances that have been officially identified as used in the manufacturing of nuclear weapons:
  - Cesium-137
  - Strontium-90

5.8.2.3.26 PRODUCT

A product name, including software and service-oriented products:

- Windows
- Cheerios
- iPhone 4S

5.8.2.3.27 PROP_MISC

A proper name that does not fall into any of the entity types specified by the other entities:

- Punic Wars in **A book on the the Punic Wars**
- World Cup in **It is called the World Cup**

5.8.2.3.28 REGION

Different regions are extracted as one of the following subtypes:

- `MAJOR`– One of the fifty states of the United States, including standard abbreviations and two-letter postal code: For example
  - California
  - Hawai'i

- Calif.

The major administrative divisions of countries, such as the provinces and territories of Canada, the administrative regions of France, and so on. For example:

- British Columbia
- Puerto Rico
- Pays de la Loire
- Guam
- Bavaria

- `MINOR`– Names of counties, prefectures, districts, or analogous geographical divisions or governmental units:
  - District of Columbia
  - Orange County

5.8.2.3.29 SOCIAL_MEDIA

Entity type for extracting entities from social media feeds. The handles (also known as `ID`) and topics are extracted as one of the following subtypes:

Note:
The `SOCIAL_MEDIA` entity type supports only Twitter feeds.

- `ID_TWITTER`–Twitter handles or `ID`s starting with "`@`", for example:
  - `@SCNblogs`
  - `@sapnoticiasbr`
  - `@sapnews`
  - `@SAP_MICROSOFT`
- `TOPIC_TWITTER`–Twitter topics starting with "`#`", for example:
  - `#SAP`
  - `#Mobility`
  - `#SAPPRESS`
  - `#SAP_projects`

5.8.2.3.30 TICKER

Company stock ticker symbols used on the stock exchange. If `TICKER` entities are found adjacent to index or market names, these are included in the span, for example:

- MSFT:NYSE

- NYSE:MSFT
- NASDAQ:MSFT
- HPQ:NASDAQ

5.8.2.3.31 TIME

Designations of hours, minutes, and seconds:

- 9:00
- 9:00 a.m.
- 9:15 pm PST

5.8.2.3.32 TIME_PERIOD

Measurements of time, and ranges of time measurements:

- 5 seconds
- 1 hour, 35 minutes
- 25 years
- 5-10 minutes
- 20-30 years
- 21st century

Time expressions:

- 8 a.m.-2 p.m.
- 2 to 5 p.m.

Date expressions:

- 2-4 May
- 3 June to 5 July

Decades, centuries, and year expressions:

- 1950s
- 50s
- 1999-2000

5.8.2.3.33 TITLE

Titles and positions, together with affiliation if available:

- President

- Secretary of State
- Director of Marketing
- United States Attorney
- Queen of England
- Microsoft CEO
- British Prime Minister

5.8.2.3.34 URI

An address on the internet, extracted as one of the following subtypes:

- `EMAIL`–Email addresses, for example:
  - dot_com@sun.com
- `IP`–IP addresses, for example:
  - 8.22.200.3
- `URL`–Internet addresses, for example:
  - http://www.netscape.com
  - www.netscape.com
  - kcbs.com

5.8.2.3.35 VEHICLE

Methods of transportation, extracted as one of the following subtypes:

- `AIR`–Air vehicles, such as airplanes and helicopters. For example:
  - Air Force One
  - Concorde
- `LAND`–Land vehicles, including the color, year, model and make of the vehicle. For example:
  - blue 1993 Volkswagen Passat
  - 1988 red Toyota Camry
- `WATER`–Water vehicles. For example:
  - USS Cole
  - USS Constitution
- `LICENSE`–Alphanumeric sequences that conform to the US and Canadian license plate formats, when preceded by a state abbreviation:
  - NY DGR-3532

  - CA 1AVC367

- `VIN`–Vehicle Identification Numbers (`VIN`) in the following format, which always includes 17 characters:
  - 1G1JF27W8GJ178227

  Each position in the `VIN` has a particular meaning, designating among other things county code, manufacturer code, equipment code, serial number, and so on.

5.8.2.3.36 WEAPON

Weapons, extracted as one of the following subtypes:

- `BIOLOGICAL`–Names of bacteria, viruses, fungi, natural toxins, and diseases that have been officially identified as used to harm humans, plants (crops), and animals, or as potential biological threats. This also extracts entities that describe the means for the dispersal of any of these weapons:
  - Anthrax
  - ricin
- `CHEMICAL`–Names of chemical substances that have been officially identified as used to harm humans, plants (crops), and animals, or as potential chemical threats. This also extracts entities that describe the means of the dispersal of any of these weapons:
  - VX
  - tabun
- `EXPLODING`–Names of substances that cause damage by exploding:
  - Molotov
  - Dynamite
- `NUCLEAR`–Names of weapons that have been officially identified as used to harm humans, plants (crops), and animals through the dispersal of radiological or nuclear energies, or have been identified as potential nuclear threats:
  - A-bomb
  - plutonium
- `PROJECTILE`–Names of weapons that are designed or used to be projected at great speed for the purpose of causing damage:
  - Stinger
  - Silkworm
- `SHOOTING`–Names of weapons that are designed or used to send projectile objects at great speed for the purpose of causing damage:
  - AK-47
  - AKM

5.8.2.3.37 YEAR

All years, including those with designators such as A.D., BC, BCE, or C.E.:

- 2001
- '63
- 1998 A.D.
- 200 BC
- 2525 C.E.

5.9 French Language Reference

This chapter describes the behavior of the French language module.

5.9.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of French texts, including word segmentation, stemming, and tagging.

5.9.1.1 Character Encodings for French

- iso_8859_1
- cp_1252
- utf_8, utf_16, ucs_4

5.9.1.2 Word Segmentation in French

The French segmenter follows all of the general segmentation rules in the white space languages. The French segmenter has the following language-specific behavior.

French clitics and elisions are separated from the words they modify. The segmenter leaves the hyphen on the end of the verb and prefixes each clitic with a hyphen. When separating elisions, the apostrophe is kept with the word whose letters were elided. Abbreviations are kept together with their punctuation.

Text	Segmented
donne-le-moi	donne-
	-le
	-moi
l'abri	l'
	abri
trad.	trad.

Related Topics

- Word Segmentation
- White Space Languages

5.9.1.3 Stemming in French

This section describes the standard stemmer, the expanded inflectional stemmer, and the inflectional stemmer guesser used for stemming in French.

5.9.1.3.1 Standard Stemmer

The French stemmer follows the general stemming rules as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. Feminine occupational nouns stem to their masculine counterparts. Proper nouns stem to themselves. This is shown in the table below.

Category	Baseform	Examples
Noun	Singular	livres -> livre; actrice -> acteur
Proper Noun	Source form	France -> France
Verb	Infinitive	connais, connaissez -> connaître
Adjective	Masculine singular	grandes, grande -> grand
Adverb	Source form	probablement -> probablement

French pronouns are stemmed in the following way. All uninflecting forms stem to themselves. Plural-only forms and all personal pronouns maintain their number information. If applicable, these pronouns are stemmed to the nominative form. All other forms stem to the masculine, singular form. This is shown in the table below:

Text	Stem
beaucoup	beaucoup
plusieurs	plusieurs
elles	ils
moi	je
lesquelles	lequel

Closed class words may be regularized or they may stem to themselves.

These word categories stem to themselves: abbreviations, acronyms, interjections, numbers and onomatopoeia forms. This is shown in the table below:

Example	Stem
par ex.	par ex.
min.	min.
UNICEF	UNICEF
15km	15km

Contracted prepositions are broken into their component parts, and these stems are returned with an equal sign in between, indicating that the stems are of equal semantic importance. If the contracted preposition occurs in a multiword units, then the final contraction is broken. This is shown in the following table:

Example	Stem
au	à=le
au moment du	au moment de=le

5.9.1.3.2 Expanded Inflectional Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text (such as email, online documents, or queries). See Expanded Inflectional Stemming for the general behavior. The specifics for French are described below.

The expanded version does not require correct capitalization and accentuation.

Case Variants

The expanded version accepts lower case letters in addition to capital letters for those words where the capitals are obligatory.

Example	Output
Paris	Paris

Example	Output
paris	Paris
IBM	IBM
ibm	IBM

Deaccented Characters

The expanded version also allows deaccented characters which replace accented ones.

Example	Output
héros	héros
heros	héros
nôtre	nôtre
notre	nôtre

Hyphenation

Hyphens in non-numeric expressions are optional in the expanded version.

Example	Output
Tiers-Monde	Tiers-Monde
TiersMonde	Tiers-Monde
est-ouest	est-ouest
estouest	est-ouest

5.9.1.3.3 Inflectional Stemmer Guesser

The inflectional stemmer guesser contains a set of morpological rules that you can apply to words that are unknown to the standard or expanded inflectional stemmer and therefore cannot be stemmed.

The software's linguistics processing first attempts to perform stemming using the standard or expanded inflectional stemmer, and then applies the stemmer guesser only to words that cannot be conventionally stemmed.

5.9.1.4 Part-of-Speech Tagging in French

The following table shows the French tag set. The tag names are accompanied by a brief description and one or more examples. If the example consists of more than one word, the word exemplifying the current tag is in **bold**. The tag set makes no distinction for gender.

Umbrella Tag	Complete Tag	Description	Examples
Adj	Adj	Invariant adjective	heureux
	Adj-Ord-Pl	Spelled out plural ordinal number	deuxièmes
	Adj-Ord-Sg	Spelled out singular ordinal number	deuxième
	Adj-Pl	Plural adjective	gentilles
	Adj-Pre	Preposed invariant adjective	vieux
	Adj-Pre-Pl	Preposed plural adjective	**petits** chiens
	Adj-Pre-Sg	Preposed singular adjective	**petit** chien
	Adj-Sg	Singular adjective	gentille
Adv	Adv	Adverb	finalement, aujourd'hui
	Adv-Deg	Adverb that can modify an adjective	très

Umbrella Tag	Complete Tag	Description	Examples
Aux	Aux-Fin-12	1st or 2nd person auxiliary, any tense	fusse
	Aux-Fin-3-Pl	3rd person plural auxiliary, any tense	seraient
	Aux-Fin-3-Sg	3rd person singular auxiliary, any tense	serait
	Aux-Inf	Infinitive auxiliary	en **avoir** assez
	Aux-PaPart	Past participle auxiliary	eu
	Aux-PrPart	Present participle auxiliary	ayant
Conj	Conj-Adv	Connecting or subordinating conjunction	quand
	Conj-comme	**Comme**	comme
	Conj-Coord	Coordinating conjunction	et, ou
	Conj-que	**Que**	que
Det	Det-Pl	Plural determiner	vos
	Det-Sg	Singular determiner	ma, votre
Interj	Interj	Interjection	hi, pouah
Nn	Nn	Invariant noun	taux
	Nn-Letter	Letter	z, K
	Nn-Net	URL and e-mail address	www.inxight.com, info@inxight.com
	Nn-Pl	Plural noun	chiens, fourmis
	Nn-Sg	Singular noun	chien, fourmi
Num	Num	Numeral or digit expression	treize, un million, 12, 15kHz, XIX
Part	Part-Neg	Negation particle	ne
	Part-voicila	Particles **voici** and **voilà**	voici, voilà

Umbrella Tag	Complete Tag	Description	Examples
Prep	Prep	Preposition (other than **à au de du** ...)	dans
	Prep-a	Preposition **à**	à, au, aux
	Prep-de	Preposition **de**	de, des, du, d'
	Prep-en	Preposition **en**	**en** bonne santé
Pron	Pron	Pronoun	il, elles
	Pron-12	1st or 2nd person pronoun	je
	Pron-Clit	Clitic pronoun	donne-**le**, donne-**lui**
	Pron-IntRel	Relative or interrogative pronoun (except **que**)	qui, quoi, lequel
Prop	Prop	Proper noun	Marie, Paris
Punct	Punct	Punctuation	: -
	Punct-Comma	Comma	,
	Punct-Quote	Quotation marks	"
	Punct-Sent	Sentence-ending punctuation	. ! ?;
V/Adj	V/Adj-PaPart	Invariant past participial verb or adjective	souri
	V/Adj-PaPart-Pl	Plural past participial verb or adjective	lues
	V/Adj-PaPart-Sg	Singular past participial verb or adjective	dansé

Umbrella Tag	Complete Tag	Description	Examples
V	V-Fin-12	1st or 2nd person verb, any tense	dansiez, dansais
	V-Fin-3-Pl	3rd person plural verb, any tense	danseront
	V-Fin-3-Sg	3rd person singular verb, any tense	dansait
	V-Inf	Infinitive verb	danser, finir
	V-PrPart	Present participle verb	notant

5.9.1.4.1 Unfound Words

Words not found in the tagger dictionary are passed to the French tagger guesser where they are assigned a tag based on a set of rules about French morphology and capitalization. The following set of tagging rules are part of this module.

Verb tags are assigned according to the verb conjugation patterns. The adverb tag is assigned to words ending in **-ement**, **-amment**, **-emment**, **-iment**. Words ending in**-able(s)**, **-ible(s)**, **-eux**, **-ois** are guessed as adjectives, and words ending in **-gé(s)**, **-ré(s)** as past participles.

Every other lowercase all-alpha word (not ending in an **-s**) is guessed as a singular noun, lowercase all-alpha words ending in **-s**, **-aux**, and **-men** are guessed as plural nouns, and lowercase all-alpha words ending in **x**, **z**, **ais**, **ois** are guessed as invariant nouns. Internet and e-mail addresses are tagged as `Nn-Net`.

Words beginning with a capital letter or a number followed by a capital letter are guessed as proper nouns. The remainder of the word may also contain numbers, lowercase or uppercase letters, hyphen or slash. Combinations of digits and punctuation are tagged as numbers. A series of punctuation marks is tagged as punctuation.

5.9.2 Extraction

This section describes the extraction-specific information for French.

5.9.2.1 French Subtypes

French supports subtypes in the types `NIN`, `ORGANIZATION`, `REGION`, `SOCIAL_MEDIA`, and `URI`.

Related Topics

- Subtypes

5.9.2.2 Predefined Entity Types

This section describes the predefined entity types supported by the French language module and examples of each. Click each link to jump to that subsection: ADDRESS1, CONTINENT, COUNTRY, CURRENCY, DATE, DAY, GEO_AREA, GEO_FEATURE, HOLIDAY, LANGUAGE, LOCALITY,MEASURE, MONTH, NIN, NOUN_GROUP, ORGANIZATION, PEOPLE, PERCENT, PERSON, PHONE, PRODUCT, PROP_MISC, REGION, SOCIAL_MEDIA, TICKER, TIME, TIME_PERIOD, TITLE,URI, and YEAR.

5.9.2.2.1 ADDRESS1

The format for `ADDRESS1` is based on typical address patterns found in Canadian French and European French addresses:

Street-address	City	Province	Country	Postal-code
27 rue Pasteur	Sherbrooke	Québec	Canada	J1K 2Y3

- 4, rue du 8 Mai 1945, Vancouver, BC, V6E 1R8
- Case Postale 123, Succursale Centre-Ville, Montréal, PQ, Canada

Street-address	Postal-code	City	Country
31bis, Saint-Joseph nord	13402	Marseille,	France

- Rue du Cornet 6 B-4800 VERVIERS BELGIQUE
- 19 quai de la Voltaire, Paris
- 68bis avenue des Abesses
- 1, av Carnot

- 2ter, Fbg des Abesses

5.9.2.2.2 CONTINENT

Any of the continents, for example:

- Asie
- Europe
- Afrique

5.9.2.2.3 COUNTRY

Names of countries, and abbreviations of a limited set of country names. This list also includes the names of geo-political entities for which conventional labels do not apply, such as disputed territories or territories that have not been internationally recognized:

- France
- République française
- Belgique
- Royaume de Belgique
- Canada
- Suisse
- É.-U.
- Gibraltar
- Kosovo
- Taiwan
- Tibet

5.9.2.2.4 CURRENCY

Expressions denoting amounts of money, and expressions denoting ranges of amounts of money:

- 6 800 000 DM
- 68.985FB
- 300,687 €
- 28,5 £
- $ 5000
- deux cent deux yens
- une dizaine de milliards de francs belges

- de 3 à 4 francs
- de 5DM à 15DM
- entre trois et quatre milliards de dollars canadiens
- 1260000000 EUR
- 0.18 EUR
- 200000 USD
- 50000000000 USD
- 36.33 USD

5.9.2.2.5 DATE

Dates are minimally composed of a number and month name:

- 6 Oct
- 10 nov.
- 17 fév 1999
- 31/12/1986
- 3-31-2000
- 3.31.2000
- 2007-07-30

Date expressions:

- 3 et 4 juin, 2000
- 3, 5, et 7 juin, 2000
- du 3 au 5 juin 2000
- 28 et 29 juillet

5.9.2.2.6 DAY

Names of the days of the week, and expression based on day names:

- vendredi
- Mardi
- jeudi et vendredi
- de samedi à dimanche
- vendredi, samedi et dimanche

5.9.2.2.7 GEO_AREA

A geographical area larger than a city that captures a significant land mass, such as a group of countries:

- Gaspésie
- Amazonie
- Asie du Sud-Est
- Silicon Valley
- Moyen-Orient
- Afrique de l'Ouest

5.9.2.2.8 GEO_FEATURE

Names of places that are not identified as `CONTINENT`, `COUNTRY`, `GEO_AREA`, `LOCALITY`, or `REGION`:

- delta du Niger
- Himalaya
- fleuve Saint-Laurent
- mer Ionienne

5.9.2.2.9 HOLIDAY

Names of popular holidays:

- Toussaint
- Nouvel An
- Pâques
- Réveillon de Noël

5.9.2.2.10 LANGUAGE

Nouns referring to languages:

- Il parle l'**espagnol**
- Un livre en **allemand**
- Le **swahili** est une langue d'Afrique

5.9.2.2.11 LOCALITY

Name of a city:

- Honfleur
- Bruxelles
- Londres
- Prague
- San Francisco

5.9.2.2.12 MEASURE

Measure expressions, and expressions denoting measure ranges:

- 200.000 tonnes
- 1.600 mégawatts
- 242.000 barils
- 45 degrés
- 18 kilomètres
- 30 ml
- 512 bits
- entre 5 et 6 centimètres
- de 50 à 60 eV

5.9.2.2.13 MONTH

Names of the months of the year, and phrases denoting more than one month:

- septembre
- mi-août
- entre avril et juin
- d'avril à juin
- mi-décembre
- de mai à septembre 1896

5.9.2.2.14 NIN

Canadian Social Insurance numbers and French INSEE numbers are extracted as one of the following subtypes:

- FR_INSEE– Numbers from the French Institut national de la statistique et des études économiques:
  - 1 23 45 67 890 000
- CA_SIN– Canadian Social Insurance numbers:

- 123-456-789

Note:
A custom Cleansing Package can be created to parse and standardize `NIN/CA_SIN` or `NIN/FR_INSEE` entities. Any extracted `NIN/CA_SIN` or `NIN/FR_INSEE` entities can also be parsed and standardized using the Data Cleanse transform of Data Services by mapping them to one of the `UDPM` (user defined pattern matching) input fields.

For details on using UDPM input fields, see *SAP Business Objects Data Services Reference Guide*

5.9.2.2.15 NOUN_GROUP

A simple noun phrase in French consists of a noun with optional premodifiers and optional postmodifers:

- progression équivalente
- développement durable
- épargne populaire
- pays européens
- internautes expérimentés
- fonds spéculatif australien

5.9.2.2.16 ORGANIZATION

Commercial, governmental, educational, legal, and service agencies, including non-profit organizations, fine arts groups, and other associations and institutions, extracted as one of the following subtypes:

- `COMMERCIAL`–The name of commercial organizations, such as major companies or corporations. For example:
  - Airbus
  - Enron
  - Northern Trust
  - Banque de Montréal
  - XYZ S.A. as in Peugeot S.A.
  - XYZ Enregistrée as in Goblet Systems Enregistrée
  - XYZ Incorporée as in Goblet Incorporée
  - Télécom XYZ as in Télécom Intermosane

  A limited number of patterns for English company names:
  - XYZ and Co. as in Arditon and Co.
  - XYZ Limited as in Advance Technology Limited

Note:
Any extracted `ORGANIZATION/COMMERCIAL` entities can be parsed and standardized using the Data Quality Data Cleanse transform by mapping them to one of the `FIRM` input fields.

- `EDUCATIONAL`–The names of institutions focused primarily on education, for example:
  - Université de Bordeaux
  - Université Pierre et Marie Curie
  - Université du Québec
  - Ecole élémentaire publique d'Amiens
- `OTHER`–Any other non-commercial organization, including groupings of geopolitical entities that can function as political entities:
  - Agence Européenne de la Sécurité Aérienne
  - Université de Louvain
  - Unesco
  - Union Européenne
  - Benelux

5.9.2.2.17 PEOPLE

Names of nationalities:

- les Anglais
- les Canadiens
- les Danois

5.9.2.2.18 PERCENT

Percent expressions, and expressions denoting measure ranges:

- 26.8%
- 6.6%
- de 70 à 85 %

5.9.2.2.19 PERSON

Variations of person names:

- François Guérard
- Clinton
- William J. Clinton

- W. J. Clinton

Full name or last name preceded by a title abbreviation:

- M. Thibaut
- Mr. Bill H. Jones
- Lieut. Van Damme
- très hon. Jean Chrétien

5.9.2.2.20 PHONE

Phone numbers based on the North American format:

- 1-800-555-1111
- (408) 555-1111
- 555-1111

Phone numbers based on the pattern used in France and internationally:

- 12 34 56 78 90
- 12/34/56/78/90
- (01) 12 34 56 78 90
- +44 (0) 1252 761314
- Tél. : +33 1 41 25 38 15
- tel +32 2 423 17 67
- Fax: +33 (0)1 55 77 33 96

5.9.2.2.21 PRODUCT

A product name, optionally preceded by a company name:

- PlayStation
- iPhone
- Airbus A320
- Boeing 737

5.9.2.2.22 PROP_MISC

A proper name that does not fall into any of the entity types specified by the other entities:

- Gemstar-TV
- EurObserver

- CeBIT
- Enduring Freedom

5.9.2.2.23 REGION

Different regions are extracted as one of the following subtypes:

- `MAJOR`– The major administrative divisions of countries, such as the provinces and territories of Canada, the administrative regions of France, and the states of the United States:
  - Alsace
  - Bretagne
  - Lorraine
  - Saint-Pierre-et-Miquelon
  - Ontario
  - Andalousie
- `MINOR`– Names of counties, prefectures, districts, or analogous geographical divisions or governmental units:
  - Calvados
  - Essonne
  - Finistère

5.9.2.2.24 SOCIAL_MEDIA

Entity type for extracting entities from social media feeds. The handles (also known as `ID`) and topics are extracted as one of the following subtypes:

Note:
The `SOCIAL_MEDIA` entity type supports only Twitter feeds.

- `ID_TWITTER`–Twitter handles or `ID`s starting with "`@`", for example:
  - `@LaurenceDutour`
  - `@René_Latendresse`
  - `@sapnoticiasbr`
  - `@SCNblogs`
  - `@sapnews`
  - `@SAP_MICROSOFT`
- `TOPIC_TWITTER`–Twitter topics starting with "`#`", for example:
  - `#Ardèche`

- `#CharliHeddo`
- `#SAP`
- `#Mobility`
- `#SAPPRESS`
- `#SAP_projects`

5.9.2.2.25 TICKER

Company stock ticker symbols used on the stock exchange:

- NYSE: SAP
- Nasdaq: BOBJ
- Nasdaq: US7170811035

5.9.2.2.26 TIME

Clock times and time expressions:

- 21h35
- 21 h 35
- 21h 35
- 21h
- 21:35
- 21:35:15
- 21H00 GMT

Clock time expressions in words:

- 3 heures
- 3 heures et quart
- midi moins un quart

Expressions based on clock times:

- entre 3h 30 et 4h
- de 12h 20 à 6h 10

5.9.2.2.27 TIME_PERIOD

Measures of time duration, and expressions denoting ranges of measures of time:

- 20 ans

- deux années
- deux jours
- dix-huit ans
- quatre-vingt-dix minutes
- de trois à deux semaines
- une journée
- deux derniers jours
- 10-15 ans
- de 3 à 8 ans
- entre 30 et 55 ans

5.9.2.2.28 TITLE

Names of important positions in government, business, and other organizations:

- directeur de service
- ministre des Finances
- gouverneur de la Banque du Canada

5.9.2.2.29 URI

An address on the internet, extracted as one of the following subtypes:

- `EMAIL`–Email addresses, including Lotus Notes addresses, for example:
  - bruno.muri@wanadoo.fr
  - cnd@media.ca
  - stephane.wallon@lecho.be
  - Dupont/BOBJ@CMP
  - CTarin/Inxight@CMP
- `IP`–IP addresses, for example:
  - 8.22.200.3
- `URL`–Internet addresses, for example:
  - Monde.fr
  - www.gensdebaignade.org
  - www.champy.ca
  - www.lactualite.com

- http://clubobs.nouvelobs.com/blogs/blogs/regereau

5.9.2.2.30 YEAR

A year identifier, decade-denoting expression, century-denoting expressions, year ranges, and expressions based on years:

- 2007
- années 60
- entre 1996 et 2006
- entre 1991 et 2000
- années 90
- 2005-2007
- 2007/2008
- années 80 et 90

5.10 German Language Reference

This chapter describes the behavior of the German language module.

5.10.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of German texts, including word segmentation and stemming.

5.10.1.1 Character Encodings for German

- iso_8859_1
- cp_1252
- utf_8, utf_16, ucs_4

5.10.1.2 Word Segmentation in German

The German segmenter follows all of the general segmentation rules in the white space languages. The German segmenter has the following language-specific behavior.

The German segmenter splits contractions at the apostrophe. A few non-contractions that include apostrophes are not split at the apostrophe, because the apostrophe is part of the word.

Text	Segmented
geht's	geht
	's
auf's	auf
	's
Maxime's	Maxime's

When a compound consists of two parts joined by a conjunction, the hyphen is not separated from the partial compound. Leading hyphens are not split off if the following word begins with a lowercase letter. However, if the following word begins with an uppercase letter, the hyphen is split off.

Text	Segmented
West- und Ostgoten	West-
	und
	Ostgoten

Text	Segmented
Silbermesser und -gabel	Silbermesser
	und
	-gabel
-West	-
	West

Abbreviations are not split off from their punctuation. Ordinal numbers are also kept together with their period.

Text	Segmented
Mrd.	Mrd.
bzgl.	bzgl.
43.	43.

Related Topics

- Word Segmentation
- White Space Languages

5.10.1.3 Stemming in German

This section describes the standard stemmer, the expanded inflectional stemmer, and the inflectional stemmer guesser used for stemming in German.

5.10.1.3.1 Standard Stemmer

The German stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below.

Category	Baseform	Examples
Noun	Nominative singular	Tischen -> Tisch; Leuten -> Leute
Verb	Infinitive	schwimmt, schwamm, geschwommen -> schwimmen
Adjective	Base form	farbigen -> farbig; vag -> vage
Adverb	Source form	ganztags -> ganztags

German pronouns are stemmed in the following way. All uninflecting forms stem to themselves. Plural-only forms and all personal pronouns maintain their number and gender information. If applicable, these pronouns are stemmed to the nominative form. All other forms stem to the singular, nominative form of the given gender (if applicable). This is shown in the table below:

Text	Stem
manch	manch
ich, meiner, mir	ich
demjenigen	dasjenige,derjenige

Uninflecting categories stem to themselves, for example, abbreviations, acronyms, numbers, conjunctions, and so on. The German stemmers support both the old and new spelling variants. If the input is the new variant, it is stemmed to the old spelling. The following table shows some examples:

Text	Stem
zahlr.	zahlr.

Text	Stem
ZDF	ZDF
Delphin, Delfin	Delphin
behende, behände	behende

Contracted prepositions are broken into their component parts, and these stems are returned with an equal sign in between, indicating that the stems are of equal semantic importance. This is shown in the following table:

Text	Stem
aufs	auf=das
beim	bei=das,bei=der
zur	zu=die

5.10.1.3.2 Expanded Inflectional Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text (such as email, online documents, or queries). See Expanded Inflectional Stemming for the general behavior. The specifics for German are described below.

The expanded version of the German module covers optional hyphenation in words with obligatory hyphens, case variation, and unaccented forms of accented characters.

Hyphenation

The expanded version accepts optional hyphenation for words which normally have an obligatory hyphen.

Example	Output
MS-DOS	MS-DOS

Example	Output
MSDOS	MS-DOS

Case Variants

The expanded version accepts lower case letters in addition to capital letters for those words where the capitals are obligatory.

Example	Output
USA	USA
usa	USA

Deaccented Characters

The expanded version of German accepts completely deaccented characters in addition to accented ones.

Example	Output
Müller	Müller
Muller	Müller

Note that this is in addition to the typewriter accented characters (for example, **ue** for **ü**), which both the standard and the expanded versions of German accept.

5.10.1.3.3 Inflectional Stemmer Guesser

The inflectional stemmer guesser contains a set of morpological rules that you can apply to words that are unknown to the standard or expanded inflectional stemmer and therefore cannot be stemmed.

Linguistics processing first attempts to perform stemming using the standard or expanded inflectional stemmer, and then applies the stemmer guesser only to words that cannot be conventionally stemmed.

5.10.1.3.4 German Compound Analysis

The German stemming module includes a comprehensive mechanism for analyzing productive compounding, capable of handling an essentially infinite number of compound words. Compounding

in German can combine various parts of speech: nouns with nouns, nouns with adjectives, verbs with nouns, and so on. Hyphenated compounds are treated like other compounds.

By default, compounds are separated into their component stems. You can use the nosplit stemming variant to keep compounds together.

The sample output shows the component stems as returned by the standard stemmer. The compound boundary is marked by #.

Example	Output
Muttertag	Mutter#Tag

The baseform of a compound element is capitalized as it would be if it stood alone, no matter where it appears in the compound. Thus noun elements have capitalized stems, even if they are not the first element in the compound. Similarly, the stem of an adjective as the first element of a compound would not be capitalized.

Compounds like **Vor-** und **Nachmittag** ("before and after noon") or **Bachkonzerte und -kantaten** ("Bach concertos and cantatas") are sometimes encountered. The hyphen is not part of the baseform.

Example	Output
Vor- und Nachmittag	vor und Nachmittag
Bachkonzerte und -kantaten	Bach#Konzert und Kantate

Noun-Noun Compounding

German noun-noun compounds often contain linking elements (**Fugenelemente**) between the main elements. For instance, when the words **Herr** ("gentleman") and **Mantel** ("coat") combine to form a compound, the linking element **-en-** is inserted between them, giving **Herrenmantel** (gentleman's coat). Linking elements do not appear in the stemmer output.

Example	Output
Herrenmantel	Herr#Mantel

The German module assumes that such linking elements can be determined based on the declension class of the noun which they follow. That is, the baseform plus the linking element usually yields a standard inflected form of the relevant word (as **Herren** is the plural of **Herr**).

Some examples of noun-noun compounds follow. Note that the module can also analyze compounds combining more than two words.

Example	Output
Lehrlingsnot	Lehrling#Not
Kinderarzt	Kind#Arzt
Kindesentführung	Kind#Entführung
Obstanbaugebiet	Obst#Anbau#Gebiet
Informatik-Konzepte	Informatik#Konzept

Nominal Final Elements

Nouns are not the only possible non-final elements of German compounds ending with nouns. The module also recognizes numerals, adjectives, adjectival participles, adverbs, verb stems, and proper nouns. As already noted, stem capitalization follows usage for single words.

Example	Output
Optimalsumme	optimal#Summe
Linkskurve	links#Kurve
Goethehaus	Goethe#Haus
Waschmaschine	waschen#Maschine

Adjectival Final Elements

An adjective can combine with other adjectives to form compound adjectives. Nouns and numerals can also be the first elements of compounds with adjectival final elements. For example:

Example	Output
blaugrün	blau#grün
hellgelb	hell#gelb
graphiktauglich	Graphik#tauglich
ausbaufähig	Ausbau#fähig
zweiblättrig	zwei#blättrig
blau-grau	blau#grau

5.10.1.3.5 Non-decompounding Stemmer

In addition to the standard compound analysis, the German language module provides a variant stemmer that does not perform de-compounding. This stemmer stems the compound as a whole, but does not split the compound into separate stems. The returned stem is always a single term; and since there is no compound boundary marker, the term cannot be broken up.

Below are examples comparing the output from the nosplit stemmer with that of the standard stemmer:

Example	Standard Output	Nosplit Output
Bildungsromane	Bildung#Roman	Bildungsroman
Männerhosen	Mann#Hose	Männerhose
Hundehütten	Hund#Hütte	Hundehütte
himmelblaue	Himmel#blau	himmelblau
Rotstiften	rot#Stift	Rotstift

Example	Standard Output	Nosplit Output
ABC-Alarm	ABC#Alarm	ABC-Alarm
Informatik-Konzepte	Informatik#Konzept	Informatik-Konzept
Bereitschafts-Dienst	Bereitschaft#Dienst	Bereitschafts-Dienst
blau-grau	blau#grau	blau-grau

Because these compounds are not split up, the capitalization remains the same as in the input example; such as, the capitalization is determined by the part-of-speech of the main word. Hyphenated words also are not separated and retain their hyphens.

5.10.1.4 Part-of-Speech Tagging in German

The following table shows the German tag set. The tag names are accompanied by a brief description and one or more examples. If the example consists of more than one word, the word exemplifying the current tag is in **bold**. The tag set makes no distinction for gender.

Umbrella Tag	Complete Tag	Description	Examples
Adj	Adj-Attr	Attributive adjective	**schwarze** Katze
	Adj-Attr-Comp	Comparative attributive adjective	**kleinere** Datei
	Adj-Attr-Ord	Spelled out ordinal numbers	dritter Mann
	Adj-Attr-Sup	Superlative attributive adjective	**schnellster** Läufer
	Adj-Dem	Demonstrative adjective	solche
	Adj-Indet	Indefinite adjective	deinige
	Adj-Pred	Predicate adjective (may also be an adverb)	schnell
	Adj-Pred-Comp	Comparative predicate adjective	besser
	Adj-Pred-Sup	Superlative predicate adjective	am **schnellsten**
Adv	Adv	Adverb	immer, zwar
	Adv-Int/Rel	Interrogative or relative adverb	wieso
	Adv-Pron	Pronominal adverb	hierfür
Aux	Aux-Fin	Finite auxiliary verb	bist
	Aux-Inf	Auxiliary verb infinitive	gebildet **haben**
	Aux-PaPart	Auxiliary verb past participle	gehabt
Cmpd	Cmpd-Left	Left compound part	zwei-, Kontakt-
Conj/Adv	Conj/Adv	Conjunction or adverb	jedoch

Umbrella Tag	Complete Tag	Description	Examples
Conj	Conj-als	Conjunction **als**	als
	Conj-Coord	Coordinating conjunction	und
	Conj-Inf	Infinitival conjunction	**um** ... zu
	Conj-Post	Correlative conjunction	weder ... **noch**
	Conj-Pre	Preposed conjunction	weder
	Conj-Subord	Subordinating conjunction	weil
	Conj-wie	Conjunction **wie**	wie
Det/Pron	Det/Pron-Quant	Quantifying determiner or pronoun	**lauter** Sachen
Det	Det-Art	Article	die, das
	Det-Dem	Demonstrative determiner	diese
	Det-Indet	Indefinite determiner	keiner
	Det-Int/Rel	Interrogative determiner or pronoun	wieviel
	Det-Poss	Possessive determiner	dein
Interj	Interj	Interjection	ach, oh
Modal	Modal-Fin	Finite modal verb	darf
	Modal-Inf	Modal verb infinitive	gehorchen **sollen**
	Modal-PaPart	Modal verb past participle	segeln **gedurft**
Nn	Nn	Noun	Tisch
	Nn-Letter	Lowercase letters with or without a period and uppercase letters	Ein **e** Ein **r.**
	Nn-Net	URL and e-mail address	www.inxight.com, info@inxight.com

Umbrella Tag	Complete Tag	Description	Examples
Num	Num	Cardinal number or date	zwei, 2.3.1999
	Num-Ord	Ordinal number	43.
Part	Part-Ant	Sentential particle	danke
	Part-Comp	Comparative particle	**viel** besser
	Part-Inf	Infinitival particle	**zu** sagen
	Part-Neg	Negation particle	nicht
	Part-Num	Numerical particle	**rund** 50 Dateien
	Part-Pos	Positive particle	**zu** schnell
	Part-Pref	Separable prefix	Er rief mich **an**.
	Part-Sup	Superlative particle	**am** besten
Prep	Prep-aus	Preposition aus	aus
	Prep-Circ	Last part of a circumposition	um Himmels **willen**
	Prep-Det	Preposition-article combination	zum
	Prep-fuer	Preposition **für**	für
	Prep-Post	Postposition	dem Haus **gegenüber**
	Prep-Pre	Preposition	mit
	Prep-pro	Preposition **pro**	pro
	Prep-von	Preposition **von**	von

Umbrella Tag	Complete Tag	Description	Examples
Pron	Pron-Dem	Demonstrative pronoun	**diese** ist besser
	Pron-Dem-Inv	Uninflected demonstrative	solch ein Erfolg
	Pron-Indet	Indefinite pronoun	niemand
	Pron-Indet-Inv	Uninflected determiner	**manch** ein Mensch
	Pron-Int/Rel	Interrogative or relative pronoun	was, wer
	Pron-Int/Rel-Inv	Uninflected interrogative or relative pronoun	was für
	Pron-Pers	Personal pronoun	ich
	Pron-Poss	Possessive pronoun	**meine** sind gelb
	Pron-Recip	Reciprocal pronoun	einander
	Pron-Refl	Reflexive pronoun	sich
	Pron-Rel	Relative pronoun	die Leute, **die** . . .
Punct	Punct	Punctuation	( )
	Punct-Comma	Comma	,
	Punct-Sent	Sentence-ending punctuation	. ? !
V	V-Fin	Finite verb	schwimmt
	V-Inf	Infinitive verb	er kann **schwimmen**
	V-Izu	**zu** infinitive	auszubilden
	V-PaPart	Past participle verb	er ist **geschwommen**

5.10.1.4.1 Unfound Words

Words not found in the tagger dictionary are passed to the German tagger guesser where they are assigned a tag based on a set of rules about German morphology and capitalization. The following set of tagging rules are part of this module.

Noun tags are assigned to words ending in a number of nominal suffixes. Verb tags are assigned to lower case words ending in **-ier** and other specified endings. Adverb tags are assigned to words ending

in **-weise**, **-ens**, and **-mal**. Words with endings like **-ig**, **-isch**, and **-los** are guessed as adjectives. Internet and e-mail addresses are tagged `Nn-Net`.

Capitalized words are guessed to be nouns. These words may contain slashes, numbers, or uppercase letters in the middle (TelCo, Tel/Fax and 3Com), but not hyphens or apostrophes. Lowercase words are guessed as adjectives or adverbs. Combinations of punctuation are guessed as punctuation. Combinations of numbers and punctuation are guessed as numbers.

5.10.2 Extraction

This section describes the extraction-specific information for German.

5.10.2.1 German Subtypes

German supports subtypes in the types `ORGANIZATION`, `REGION`, `SOCIAL_MEDIA`, and `URI`.

Related Topics

- Subtypes

5.10.2.2 Predefined Entity Types

This section describes the predefined entity types supported by the German language module and examples of each. Click each link to jump to that subsection: ADDRESS1, CONTINENT, COUNTRY, CURRENCY, DATE, DAY, GEO_AREA, GEO_FEATURE, HOLIDAY, LANGUAGE, LOCALITY, MEASURE, MONTH, NOUN_GROUP, ORGANIZATION, PEOPLE, PERCENT, PERSON, PHONE, PRODUCT, PROP_MISC, PUBLICATION, REGION, SOCIAL_MEDIA, TIME, TIME_PERIOD, TITLE, URI, and YEAR.

5.10.2.2.1 ADDRESS1

The format for addresses is based on the German address. For street addresses:

Street	Street No.	City-Code	City	Country
Kaiserstraße	123	D-10623	Berlin	Deutschland

- Kaiserstraße 123, 10623 Berlin, Deutschland
- Stockerauerstraße 9, A8700 Leoben
- Zugerbergstrasse 18, CH-6414 Unteraegeri

For post office box addresses:

P.O. Box number	City-Code	City	Country
Postfach 10 43 51	D-70049	Stuttgart	Deutschland

5.10.2.2.2 CONTINENT

Any of the continents, for example:

- Afrika
- Europa

5.10.2.2.3 COUNTRY

Names of countries and abbreviated name of the countries. This list also includes names of geo-political entities for which conventional labels do not apply, such as disputed territories or territories that have not been internationally recognized:

- Deutschland
- Vereinigte Staaten von Amerika
- U.K.
- Palästina
- Taiwan

5.10.2.2.4 CURRENCY

Quantities of world currency:

- 3$
- 85,00 DM

- DM 48,00
- US$ 1.00
- drei Dollar
- 15,- DM
- Euro 14.78
- 40 Millionen Euro

Ranges of amounts of currency:

- zwischen 3 und 4 Euro
- zwischen $3 und $4
- von 3 bis 4 Euro

Currency phrases:

- 3 Millionen Euro Umsatz
- EUR 3.000 Festgeld

Currency ratios:

- 4500 Euro pro Quadratmeter
- 119 EUR pro Aktie
- $20 / Stunde
- 0,5 Pf/min.

5.10.2.2.5 DATE

Dates:

- Montag, den 30. September 1954
- Mi., 10. November 1998
- 14. Februar 1999
- 14. Februar '99
- 7. Januar
- 16.01.2000
- 16.01.99
- 7-2-2000

Dates that span several days:

- 29.3.-2.5.2000

- 9.-10. Mai 2000
- vom 1. März bis 2. Juni
- 14. August - 15. September

5.10.2.2.6 DAY

Day of the week:

- Montag
- Mittwoch
- Mo.
- Di.

5.10.2.2.7 GEO_AREA

A geographical area larger than a city that captures a significant land mass such as a group of countries:

- Südamerika
- Nordamerika
- Karibik
- Westeuropa

5.10.2.2.8 GEO_FEATURE

A place name that is not identified as a `COUNTRY`, `GEO_AREA`, `LOCALITY`, or `REGION`.

- Bodensee
- Zugspitze
- Uranus
- Indischer Ozean

5.10.2.2.9 HOLIDAY

Names of popular holidays:

- Weihnachten
- Tag der Arbeit
- Buß- und Bettag

5.10.2.2.10 LANGUAGE

Noun referring to a language

- Englisch
- Deutsch
- Portugiesisch

5.10.2.2.11 LOCALITY

City names:

- Paris
- San Francisco
- La Paz

City names with geographical specification:

- Freiburg im Breisgau
- Frankfurt am Main
- Frankfurt a.d. Oder

5.10.2.2.12 MEASURE

Measure expressions:

- 2800 Angström
- 50 eV
- 58,68 mm
- 9kWh
- 25 Grad Celsius
- 90° Fahrenheit
- sechshundert Kilogramm

Measure Ranges:

- von 50 bis 60 Kilometer
- von 10 Volt bis 20 Volt
- zwischen 5 und 6 Zentimetern

Measure ratios:

- 5 km/h
- 33g/l
- 27 mg/Kubikmeter
- 5 Kilometer pro Stunde

5.10.2.2.13 MONTH

Names of the months of the year:

- Januar
- Dezember
- Jan.
- Mrz.

5.10.2.2.14 NOUN_GROUP

German noun phrases include hyphenated nouns and adjective-noun groups:

- moderne Technologien
- fachliches Wissen
- rationelle Terminplanung
- Ingenieur-Kompetenz

5.10.2.2.15 ORGANIZATION

Government, legal, and service agencies, including non-profit organizations, fine arts groups, and other associations and institutions, extracted as one of the following subtypes:

- `COMMERCIAL`–The name of commercial organizations, such as major companies or corporations. For example:
  - Mercedes Benz
  - Siemens
  - Dr. Nacken und Partner
  - Texas Instruments
  - Siemens AG
  - Otto Wolff Kunststoffvertrieb GmbH
  - Bartsch und Partner GbR
  - Ormecon Chemie GmbH & Co., KG
  - Walter de Gruyter, Inc.
  - Greening Donald Co. Ltd.
  - Volksbank Hamburg
  - Volksbank Bonn Rhein-Sieg
  - Dresdner Raiffeisenbank eG

Note:
Any extracted `ORGANIZATION/COMMERCIAL` entities can be parsed and standardized using the Data Quality Data Cleanse transform by mapping them to one of the `FIRM` input fields.

- `EDUCATIONAL`–The names of institutions focused primarily on education, for example:
  - Freie Universität Berlin
  - Rheinisch-Westfälische Technische Hochschule Aachen
- `OTHER`–Any organization that does not fit into a more specific subtype including groupings of geopolitical entities that can function as political entities:
  - Die Grünen
  - Landesamt für Statistik
  - Greenpeace
  - Sozialdemokratische Partei Deutschlands
  - Zentralstelle für Agrardokumentation und -information
  - Benelux
  - Commonwealth

5.10.2.2.16 PEOPLE

Name referring to a group of people based on country, ethnicity, or region

- Schwabe
- Amerikanerin

5.10.2.2.17 PERCENT

Percent expressions:

- 1%
- + 1,234%
- 3 Prozent
- drei Prozent
- 4 1/2 Prozentpunkte

Percent ratios and ranges:

- 7% pro Jahr
- 5%/Jahr
- 34%/Person
- 2 bis 3 Prozent

- 2-3%
- 1-1,5 Prozent

5.10.2.2.18 PERSON

Variations of names:

- Maria Hildebrandt
- Hans Peter Mayer
- Heidi
- Anne-Marie
- Vera F. Burkhardt
- Otto von Gruber

5.10.2.2.19 PHONE

German, Austrian and Swiss phone numbers:

- +49 1111 2222
- (+49)-111-22-33333
- Telefon 01 11/11 22 33
- Fax: 0111 - 22222
- Tel. 01111-1111
- T 030/22 22 200

5.10.2.2.20 PRODUCT

Product names:

- Ford Explorer
- Windows
- Jacobs Kaffee

5.10.2.2.21 PROP_MISC

One or more proper nouns in a row; names of events such as exhibitions or sports events:

- CeBIT
- Internationale Funkausstellung
- Europacup

- Olympiade

5.10.2.2.22 PUBLICATION

A newspaper, magazine or journal:

- Handelsblatt
- Frankfurter Allgemeine Zeitung
- Fürther Nachrichten

5.10.2.2.23 REGION

Different regions are extracted as one of the following subtypes:

- `MAJOR`– The major administrative divisions of countries, such as the the administrative regions of Germany, and the states of the United States: For example
  - German Bundesländer:
    - Hessen
    - Baden-Württemberg
    - Sachsen-Anhalt

  States from other countries:
  - California
  - New York State
  - British Columbia
- `MINOR`– Names of counties, prefectures, districts, or analogous geographical divisions or governmental units:
  - Mittelfranken
  - Landkreis Pfaffenhofen
  - Kreis Kelheim

5.10.2.2.24 SOCIAL_MEDIA

Entity type for extracting entities from social media feeds. The handles (also known as `ID`) and topics are extracted as one of the following subtypes:

Note:
The `SOCIAL_MEDIA` entity type supports only Twitter feeds.

- `ID_TWITTER`–Twitter handles or `ID`s starting with "`@`", for example:
  - `@Matthias_123`

    - `@RüdigerSchmitz`
    - `@SCNblogs`
    - `@sapnoticiasbr`
    - `@sapnews`
    - `@SAP_MICROSOFT`
  - `TOPIC_TWITTER`–Twitter topics starting with "`#`", for example:
    - `#Griechenland`
    - `#Mobility`
    - `#SAP`
    - `#SAPPRESS`
    - `#SAP_projects`
    - `#Weihnachten`

5.10.2.2.25 TIME

Time expressions:

- 18:05:48
- 02:00 MET
- 16.15 Uhr
- 2h 39
- 16h 45

5.10.2.2.26 TIME_PERIOD

Measures of time duration, and expressions denoting ranges of time:

- 27 Jahre
- 0,6 sec.
- 9 - 12 Monate
- im 1. Halbjahr '99
- des ersten Quartals 1999

5.10.2.2.27 TITLE

An individual specified by title only, no name:

- Schah

- Papst
- Königin
- Dr.
- Professor
- Bundesgesundheitsministerin

5.10.2.2.28 URI

An address on the internet, extracted as one of the following subtypes:

- `EMAIL`–Email addresses, for example:
  - dot_com@sun.com
- `IP`–IP addresses, for example:
  - 8.22.200.3
- `URL`–Internet addresses, for example:
  - `IP`–IP adresses, for example:
    - 8.22.200.3
  - http://www.netscape.com
  - www.netscape.com
  - kcbs.com

5.10.2.2.29 YEAR

A year identifier, decade-denoting expression, century-denoting expressions, year ranges, and expressions based on years:

- '99
- 58 vor Christus
- 200 v.Chr.
- 3 n. Chr.
- 11. Jhdt.
- 6. Jh. vor Christus
- 2000-1
- 404-399 v.Chr.
- in den Jahren 1488 bis 1490

5.11 Greek Language Reference

This chapter describes the behavior of the Greek language module.

5.11.1 Linguistic Processing

This section describes the language-specific information on the processing of Greek texts, including word segmentation and stemming.

5.11.1.1 Character Encodings for Greek

- iso_8859_7
- cp_1253
- utf_8, utf_16, ucs_4

5.11.1.2 Word Segmentation in Greek

The Greek segmenter follows all of the general segmentation rules in the white space languages.

Related Topics

- Word Segmentation
- White Space Languages

5.11.1.3 Stemming in Greek

The Greek stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Examples
Noun	αποζημιώσεις -> αποζημίωση, όζαινες -> όζαινα, ευμάρειαν -> ευμάρεια
Verb	κοροϊδεμένη -> κοροϊδεύω, φουχτωθούμε -> φουχτώνω, μισθοδοτημένους -> μισθοδοτώ
Adjective	αρθρωτήν -> αρθρωτός, πλαστικές -> πλαστικός, μονόσπερμοι -> μονόσπερμος
Adverb	πόθεν -> πόθεν, κατανυκτικότερα -> κατανυκτικά

5.11.2 Extraction

Note:
Greek is a basic-level supported language, which means it supports extraction by using dictionaries or extraction rules only.

5.12 Hungarian Language Reference

This chapter describes the behavior of the Hungarian language module.

5.12.1 Linguistic Processing

This section describes the language-specific information on the processing of Hungarian texts, including word segmentation and stemming.

5.12.1.1 Character Encodings for Hungarian

- iso_8859_2
- cp_1250
- utf_8, utf_16, ucs_4

5.12.1.2 Word Segmentation in Hungarian

The Hungarian segmenter follows all of the general segmentation rules in the white space languages.

Related Topics

- Word Segmentation

5.12.1.3 Stemming in Hungarian

Stemming in Hungarian includes the standard stemmer and the expanded stemmer.

5.12.1.3.1 Standard Stemmer

The Hungarian stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Examples
Noun	hírt -> hír, vállalatnak -> vállalat, bevételei -> bevétel
Verb	kérek -> kér, ünnepelnek -> ünnepel, élünk -> él
Adjective	privatizációssal -> privatizációs, frisset -> friss, japánok -> japán

Category	Examples
Adverb	már -> már, majd -> majd

5.12.1.3.2 Expanded Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text, such as email, online documents, or queries. In Hungarian, this includes accented characters missing their diacritics and proper names without word-initial capitalization. For instance:

Case Variants

The expanded version accepts lower case letters in addition to capital letters for those words where the capitals are obligatory.

Example	Output
Júnót	Júnó
junot	Júnó

Deaccented Characters

The expanded version also allows deaccented characters in place of accented ones.

Example	Output
tunulnék	tanul
tunulnek	tanul
Junot	Júnó

5.12.2 Extraction

Note:
Hungarian is a basic-level supported language, which means it supports extraction by using dictionaries or extraction rules only.

5.13 Italian Language Reference

This chapter describes the behavior of the Italian language module.

5.13.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Italian texts, including word segmentation and stemming.

5.13.1.1 Character Encodings for Italian

- iso_8859_1
- cp_1252
- utf_8, utf_16, ucs_4

5.13.1.2 Word Segmentation in Italian

The Italian segmenter follows all of the general segmentation rules in the white space languages. The Italian segmenter has the following language-specific behavior.

The segmenter separates Italian elisions, including elided numbers, from the words they modify. When separating elisions, the apostrophe is kept with the word whose letters were elided. Combined words that are written without apostrophes are not split.

Text	Segmented
d'un'artistica	d'
	un'
	artistica
cinqu'inviati	cinqu'
	inviati
nella	nella

Related Topics

- Word Segmentation

5.13.1.3 Stemming in Italian

This section describes the standard stemmer and the expanded stemmer used for stemming in Italian.

5.13.1.3.1 Standard Stemmer

The Italian stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Baseform	Examples
Noun	Singular	capi -> capo, pagine -> pagina
Verb	Infinitive	andiamo -> andare; parlava -> parlare
Adjective	Masculine singular	alte -> alto; grandissimo -> grande

Category	Baseform	Examples
Adverb	Source form	contentamente -> contenta-mente; più -> più

Contracted prepositions and pronouns are broken into their component parts, and these stems are returned with an equal sign in between, indicating that the stems are of equal semantic importance. This is shown in the following table:

Example	Stem
allo	a=lo
d'una	di=uno
glielo	lui=lui

5.13.1.3.2 Expanded Inflectional Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text (such as email, online documents, or queries). See Expanded Inflectional Stemming for general behavior. The specifics for Italian follow.

The expanded version does not require correct capitalization and accentuation.

Case Variants

The expanded version accepts lower case letters in addition to capital letters for those words where the capitals are obligatory.

Example	Output
Roma	Roma
roma	Roma
USA	USA

Example	Output
usa	USA

Deaccented Characters

The expanded version also allows deaccented characters in place of accented ones.

Example	Output
città	città
citta	città

Hyphenation

Hyphens in non-numeric expressions are optional in the expanded version.

Example	Output
Clermont-Ferrand	Clermont-Ferrand
ClermontFerrand	Clermont-Ferrand
liberal-democratico	liberal-democratico
liberaldemocratico	liberal-democratico

5.13.1.3.3 Inflectional Stemmer Guesser

The inflectional stemmer guesser contains a set of morpological rules that you can apply to words that are unknown to the standard or expanded inflectional stemmer and therefore cannot be stemmed. You can employ the stemmer guesser by combining it with a standard or expanded inflectional stemmer for the corresponding language, in which case the inflectional guesser should be used as the last stemmer. This enables you to have linguistics processing first attempt to perform stemming using the standard or expanded inflectional stemmer, and then apply the stemmer guesser only to words that cannot be conventionally stemmed.

5.13.1.4 Part-of-Speech Tagging in Italian

The following table shows the Italian tag set available for using in defining custom entities. The tag names are accompanied by a brief description and one or more examples. If the example consists of more than one word, the word exemplifying the current tag is in **bold**. The tag set makes no distinction for gender.

Umbrella Tag	Complete Tag	Description	Examples
Adj	Adj-ex	The prefix **ex**	ex
	Adj-Pl	Plural adjective (includes ordinals: **secondi**)	belle
	Adj-PrPart-Pl	Plural present participle verb	meditanti, destreggianti
	Adj-PrPart-Pl-Pron	Plural present participle verb with attached clitic	fasciantemela, quietanteti
	Adj-PrPart-Sg	Singular present participle verb	meditante, destreggiante
	Adj-PrPart-Sg-Pron	Singular present participle verb with attached clitic	epurantelo, andantevi
	Adj-Sg	Singular adjective (includes ordinals: **secondo**, **2º**)	buono, narcisistico
Adv	Adv	Adverb	fumettisticamente

Umbrella Tag	Complete Tag	Description	Examples
Aux	Aux	Finite auxiliary (be and have)	saranno, avrete
	Aux-Ger	Gerundive auxiliary	essendo, avendo
	Aux-Impv	Imperative auxiliary	sii, abbi
	Aux-Inf	Infinitive auxiliary	esser, aver
	Aux-PaPart-Pl	Plural past participle auxiliary	avuti, avute
	Aux-PaPart-Sg	Singular past participle auxiliary	avuta, avuto
	Aux-PrPart-Pl	Plural present participle auxiliary	essenti, aventi
	Aux-PrPart-Sg	Singular present participle auxiliary	essente, avente
Conj	Conj	Conjunction	tuttavia
	Conj-Adv	Interrogative adverb	quando, dove, come
	Conj-che	The connector **che**	ch', che
	Conj-Coord	Coordinating conjunction	ed, e/o
	Conj-Pre	First part of a multiword conjunction	**dato** che

Umbrella Tag	Complete Tag	Description	Examples
Det/Pron	Det/Pron-Int-Pl	Plural interrogative determiner or pronoun	**quanti** soldi
	Det/Pron-Int-Sg	Singular interrogative determiner or pronoun	qual, cos'
	Det/Pron-Poss-Pl	Plural possessive determiner or pronoun	mie, vostri
	Det/Pron-Poss-Sg	Singular possessive determiner or pronoun	nostro, sua
	Det/Pron-Quant	Invariant quantifying determiner or pronoun	qualunque, qualsivoglia
	Det/Pron-Quant-Pl	Plural quantifying determiner or pronoun	**molti** uomini
	Det/Pron-Quant-Sg	Singular quantifying determiner or pronoun	**molta** gente
Det	Det-Pl	Plural determiner	quei
	Det-Pre	Pre-determiner	**tutto** il giorno
	Det-Sg	Singular determiner	quel
Interj	Interj	Interjection or onomatopoeia	uhi, perdiana, eh
Nn	Nn-Letter	Lowercase and uppercase letters, by themselves or followed by a period or right parenthesis	b, N
	Nn-Net	URL, e-mail address	www.inxight.com, info@inxight.com
	Nn-Pl	Plural noun	case
	Nn-Sg	Singular noun	casa, balsamo
Num	Num	Numeric expression (in digits)	+5, 23.05, 3,45, 1997

Umbrella Tag	Complete Tag	Description	Examples
Prep	Prep	Preposition	tra, con
	Prep-a	Preposition **a**	a
	Prep-da	Preposition **da**	da
	Prep-Det-Pl	Combination preposition and plural determiner	sulle, sugl', pegli
	Prep-Det-Pl-a	Combination **a** and plural determiner	ai, alle
	Prep-Det-Pl-da	Combination **da** and plural determiner	dalle
	Prep-Det-Pl-di	Combination **di** and plural determiner	delle
	Prep-Det-Sg	Combination preposition and singular determiner	sullo, nella
	Prep-Det-Sg-a	Combination a and sg. determiner	al, allo
	Prep-Det-Sg-da	Combination **da** and singular determiner	dalla
	Prep-Det-Sg-di	Combination **di** and singular determiner	delle
	Prep-di	Preposition **di**	di
	Prep-Pre	First word of a multi-word preposition	per mezzo

Umbrella Tag	Complete Tag	Description	Examples
Pron	Pron	Invariant pronoun	sé
	Pron-chi	Interrogative pronoun **chi**	chi
	Pron-Clitic	Clitic pronoun	vi, ne, mi, glielo
	Pron-Clitic-Pre	First of a two-clitic sequence	ce, ve
	Pron-Indef-Pl	Plural indefinite pronoun	**Tutti** amano le vacaze.
	Pron-Indef-Sg	Singular indefinite pronoun	qualcuno
	Pron-Pl	Plural pronoun	noi
	Pron-Rel	Invariant relative pronoun	cui
	Pron-Rel-Pl	Plural relative pronoun	i bambini i **quali**
	Pron-Rel-Sg	Singular relative pronoun	il bambino il **quale**
	Pron-Sg	Singular pronoun	lei, lui
Prop	Prop	Proper noun	Bernardo, Monte Isola
Punct	Punct	Punctuation	: - \
	Punct-Comma	Comma	,
	Punct-Sent	Sentence punctuation	. ! ? ;
V/Adj	V/Adj-PaPart-Pl	Plural past participle verb or adjective	riposti, offuscate
	V/Adj-PaPart-Pl-Pron	Plural past participle verb or adjective, with attached clitic	telatesele, assestatici
	V/Adj-PaPart-Sg	Singular past participle verb or adjective	sbudellata
	V/Adj-PaPart-Sg-Pron	Singular past participle verb or adjective, with attached clitic	commossosi, ingranditomi

Umbrella Tag	Complete Tag	Description	Examples
V	V-Fin	Finite verb	blatereremo
	V-Fin-Pron	Finite verb with attached clitic	trattansi, leggevansi
	V-Ger	Gerund	adducendo, intervistando
	V-Ger-Pron	Gerund with attached clitic	saziandotele, appurandolo
	V-Impv	Imperative verb	**Va'** a casa!
	V-Impv-Pron	Imperative verb with attached clitic	russateli, planaci
	V-Inf	Infinitive verb	sciupare, trascinar
	V-Inf-Pron	Infinitive verb with attached clitic	spulciarsi, risucchiarsi

5.13.2 Extraction

This section describes the extraction-specific information for Italian.

5.13.2.1 Italian Subtypes

Italian supports subtypes in the types NIN, ORGANIZATION, REGION, SOCIAL_MEDIA, and URI.

5.13.2.2 Predefined Entity Types

This section describes the predefined entity types supported by the Italian language module and examples of each. Click any link to jump to that subsection: ADDRESS1 and ADDRESS2, CONTINENT, COUNTRY, CURRENCY, DATE, DAY, GEO_AREA, GEO_FEATURE, HOLIDAY, LANGUAGE, LOCALITY, MEASURE, MONTH, NIN, NOUN_GROUP, ORGANIZATION, PEOPLE, PERCENT, PERSON,

PHONE, PRODUCT, PROP_MISC, REGION, SOCIAL_MEDIA, TICKER, TIME, TIME_PERIOD, TITLE, URI, and YEAR.

5.13.2.2.1 ADDRESS1 and ADDRESS2

The format of ADDRESS1 and ADDRESS2 is based on Italian addresses, which are typically of the form:

Street-address Postal-code City Province Country

ADDRESS1 corresponds to the first part of the address, the "street-address". ADDRESS2 is meant to capture the second part of the address, the "postal-code city province" sequence. Country names are always extracted as COUNTRY even in address context.

Examples of ADDRESS1:

- Via Generale Guison 1
- Casella postale 109
- Piazza Colonna, 370
- Località Vallericca, 375

Examples of ADDRESS2:

- 00144 ROMA RM
- 40141 Bologna (BO)
- 6500 Bellinzona
- Sarego (VI)
- 84070 - Stella Cilento - Salerno

5.13.2.2.2 CONTINENT

Any of the continents, for example:

- Asia
- Europe
- Africa

Note:
Australia is extracted as `COUNTRY` only.

5.13.2.2.3 COUNTRY

Names of countries and their abbreviations. This list also includes the names of geo-political entities for which conventional labels do not apply, such as disputed territories or territories that have not been internationally recognized:

- Italia
- Germania
- U.K.
- USA

- Autorità Nazionale Palestinese
- Taiwan

5.13.2.2.4 CURRENCY

Expressions denoting amounts of money, and expressions denoting ranges of amounts of money

- € 1.000,00-1.800,00
- 28,5 £
- $ 5000
- 35 centesimi
- 1.19 dlrs
- un euro e venticinque centesimi
- da 100 a 250 euro
- 1260000000 EUR

5.13.2.2.5 DATE

Dates are minimally composed of a number and month:

- 25 Aprile
- 15 Novembre 1998
- dieci Settembre
- quattro di Giugno
- 10 nov.
- 04/12/2012

5.13.2.2.6 DAY

Days of the week, including abbreviations:

- lunedì
- Lun
- GIO

5.13.2.2.7 GEO_AREA

A geographical area larger than a city that captures a significant land mass, such as a group of countries:

- Midwest
- Sudest Asiatico
- Europa dell'Est

- Boemia

5.13.2.2.8 GEO_FEATURE

A geographical location that does not constitute a political entity and is not a CONTINENT, GEO_AREA, COUNTRY, REGION, or LOCALITY:

- Tropico del Cancro
- Plutone
- Vesuvio
- Tevere
- Monte Bianco
- Adriatico
- Himalaya

5.13.2.2.9 HOLIDAY

Holidays when banks or businesses are closed (i.e., bank holidays) in the countries where Italian is the official language. Spelled out dates that are also names of holidays are extracted as HOLIDAY (e.g., "Venticinque Aprile") whereas numeric dates that coincide with a holiday are extracted as DATE (e.g., "25 Aprile").

- Venticinque Aprile
- Pasqua
- Natale
- Festa della Repubblica italiana

5.13.2.2.10 LANGUAGE

Nouns referring to languages:

- Un libro in **Arabo**.
- Lui parla **spagnolo**.
- Lo **Swahili** è una lingua bantu

5.13.2.2.11 LOCALITY

Name of a city:

- Roma
- Parigi
- New York
- Bonn

5.13.2.2.12 MEASURE

Measure expressions, and expressions denoting measure ranges:

- 25 mq
- 20 grammi

- 6m
- 200 tonnellate
- 16GB
- 65 mph
- 33 mpg
- 0,12 euro per azione
- 400 euro al mese

5.13.2.2.13 MONTH

Months of the year, including abbreviations:

- Gennaio
- Feb.
- OTT

5.13.2.2.14 NIN

Italian fiscal codes, without spaces, and Swiss AVS numbers are extracted as one of the following subtypes:

- IT_CF – Codice Fiscale
  - BNCMRA70A20H501B
- CH_AVS –Assicurazione Vecchiaia e Superstiti
  - 756.9217.0769.85

Note:
A custom Cleansing Package can be created to parse and standardize `NIN/IT_CF` or `NIN/IT-_AVS` entities. Any extracted `NIN/IT_CF` or `NIN/IT_AVS` entities can also be parsed and standardized using the Data Cleanse transform of Data Services by mapping them to one of the `UDPM` (user defined pattern matching) input fields.

For details on using UDPM input fields, see *SAP Business Objects Data Services Reference Guide*

5.13.2.2.15 NOUN_GROUP

A simple noun phrase in Italian consists of a noun with at least one premodifier or postmodifer:

- scienziato italiano
- importante gruppo finanziario
- grande metropoli

5.13.2.2.16 ORGANIZATION

Commercial, governmental, educational, legal, and service agencies, including non-profit organizations, fine arts groups, and other associations and institutions, extracted as one of the following subtypes:

- `COMMERCIAL`–The name of commercial organizations, such as major companies or corporations. For example:
  - Fiat
  - Fininvest
  - Apple Corporation
  - Peugeot S.A.
- `EDUCATIONAL`–The names of institutions focused primarily on education. For example:
  - La Sapienza
  - La Normale di Pisa
  - Harvard University
  - Università degli Studi di Firenze
- `OTHER`– Any other non-commercial and non-educational organization, including groupings of geopolitical entities that can function as political entities:
  - UNESCO
  - Unione Europea
  - Associated Press
  - Governo Monti
  - Juventus Football Club

5.13.2.2.17 PEOPLE

Names of nationalities:

- Italiani
- Svizzeri
- Americani

5.13.2.2.18 PERCENT

Percent expressions, and expressions denoting ranges of percentages:

- 220%
- cinquanta per cento
- dal 10% al 20%
- tra il 5 e il 10 per cento

5.13.2.2.19 PERSON

An individual specified by name. A variety of forms are identified:

- Mario Monti

- Barack Obama
- Angela Merkel
- William J. Clinton
- Giulio Terzi di Sant'Agata
- Sig.ra Badano
- Ing. Paolo Rossi

Note:

Given and family names that occur by themselves are extracted as PERSON as long as they are not ambiguous with common names, with the exception of famous cases, such as Monti.

5.13.2.2.20 PHONE

Phone numbers based on the pattern used in Italy and internationally:

- +39 (06) 322 0404
- +41 91 985 88 55
- Tel. : +39 06 421111
- +41 91 993 10 11 (Fax)
- 12/34/56/78/90
- (01) 12 34 56 78 90
- 1-800-555-1111
- (408) 555-1111
- 555-1111

5.13.2.2.21 PRODUCT

A product name, optionally preceded by a company name:

- PlayStation
- iPhone 4S
- Windows 7

5.13.2.2.22 PROP_MISC

A proper name that does not fall into any of the entity types specified by the other entities:

- Aeroporto Malpensa
- NASDAQ-100
- Enduring Freedom

5.13.2.2.23 REGION

Different regions are extracted as one of the following subtypes:

- MAJOR - The major administrative divisions of countries, such as the regions of Italy, the cantons of Switzerland, and the states of the United States:

  - Valle d'Aosta
  - Lazio
  - Campania
  - Canton Ticino
  - Nevada
  - Andalusia
- MINOR - Names of counties, prefectures, districts, or analogous geographical divisions or governmental units:
  - le province di **Caserta**, **Napoli**, **Roma** e **Frosinone**

5.13.2.2.24 SOCIAL_MEDIA

Entity type for extracting entities from social media feeds. The handles (also known as ID) and topics are extracted as one of the following subtypes:

- ID_TWITTER–Twitter handles or IDs starting with "@", for example:
  - @pincopallino
  - @sapnews
  - @SAP_MICROSOFT
- TOPIC_TWITTER–Twitter topics starting with "#", for example:
  - #pippopluto
  - #SAP
  - #SAP_projects

5.13.2.2.25 TICKER

Company stock ticker symbols used on the stock exchange:

- NYSE: SAP
- Nasdaq: BOBJ
- NASDAQ:AAPL

5.13.2.2.26 TIME

Clock times and time expressions:

- 8:00
- 21.30
- 12:00 a.m.
- mezzogiorno
- ore 3 e 15
- 3 e un quarto

5.13.2.2.27 TIME_PERIOD

Measures of time duration, and expressions denoting ranges of measures of time:

- 20 anni
- 2 giorni
- due anni
- quattordici minuti
- 10-15 anni
- dai 3 agli 8 minuti
- tra le 8:00 e le 15:30
- 3-4 Giugno 2000
- dal 5 al 12 Ottobre
- 2 secoli
- anni settanta
- 2000-12

5.13.2.2.28 TITLE

Names of important positions in government, business, and other organizations:

- Presidente della Arnoldo Mondadori Editore S.p.A.
- ministro degli esteri
- governatore della banca d'Italia

5.13.2.2.29 URI

An address on the internet, extracted as one of the following subtypes:

- EMAIL - Email addresses, including Lotus Notes addresses, for example:
  - peter.schmitz@sap.com
  - john dot smith at mit dot edu
  - luigi.rossi(chiocciola)it.ey.com
  - fedele[dot]marco[chiocciola]alitalia[dot]it
  - Mary Thomas/Corp/Enron@CMP
  - cidha@integrazionecoop.org
- IP - IP addresses, for example:
  - 8.22.200.3
  - 156.138.5.11
- URL - Internet addresses, for example:
  - Repubblica.it
  - http://www.sap.com
  - www.amazon.it
  - ftp://ftp.liberliber.it/liberliber/ftp/readme.txt

5.13.2.2.30 YEAR

A year identifier:

- 2012
- '48
- 300 A.C.

5.14 Japanese Language Reference

This chapter describes the behavior of the Japanese language module.

5.14.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Japanese texts, including word segmentation and stemming.

5.14.1.1 Character Encodings for Japanese

- shift_jis
- euc_jp
- utf_8, utf_16, ucs_4

5.14.1.2 Word Segmentation in Japanese

The Japanese segmenter follows all of the general segmentation rules in the non-white space languages. The Japanese segmenter has the following language-specific behavior.

The word segmenter breaks text into minimum syntactic units, called "bunsetu" in Japanese. A Japanese word, or syntactic unit, contains a number of dependent words such as case markers (for example, が, を, は) and inflectional parts of the predicate (for example, ます, た, させる).

Case markers are separated from nominal elements:

Text	Segmented
太郎が本を読む	太郎
	が
	本
	を
	読む

Inflectional suffixes are segmented together with the head verb:

Text	Segmented
来ました	来ました
食べさせられました	食べさせられました

Aspectual and modal verbs are separated from the head verb:

Text	Segmented
食べるようだ	食べる
	よう
	だ
食べはじめた	食べ
	はじめた

The copula is separated from the head noun:

Text	Segmented
本だ	本
	だ

Classifiers are attached to the preceding numeral:

Text	Segmented
３冊	３冊
三人	三人

Noun prefixes are separated from the nouns, while the verb and adjectival prefixes are attached to their heads:

Text	Segmented
高品質	高
	品質
お座り	お座り
バカでかい	バカ
	でかい

Punctuation marks, including opening and closing marks, are segmented separately:

Text	Segmented
「紅花」	「
	紅花
	」
★注意	★
	注意

The Japanese segmenters treat whitespace within hiragana and katakana sequences as syntactic unit boundaries. That is, syntactic units are broken as follows:

Text	Segmented
オフィス ソリューション	オフィス
	ソリューション

Newline characters, such as "`/n`", are preserved when surrounded by katakana words.

Hyphens and slashes also break syntactic units. Hyphenated kanji words and katakana words are separated:

Text	Segmented
東京–箱根	東京
	–
	箱根

Text	Segmented
パリ–ロンドン	パリ
	–
	ロンドン
パリ / ロンドン	パリ
	/
	ロンドン

Numeric expressions with or without punctuation marks are kept together:

Text	Segmented
12,000	12,000
20/20	20/20
25%	25%
２.５	２.５
二五–五十	二五–五十

Note:
Japanese words written entirely in hiragana script, rather than in the more standard combination of kanji and hiragana, may not be properly segmented due to ambiguity. Such written style is usually restricted to text intended for children or learners of Japanese.

Related Topics

- Word Segmentation
- White Space Languages

5.14.1.3 Stemming in Japanese

This section describes the standard stemmer and the expanded stemmer used for stemming in Japanese.

5.14.1.3.1 Standard Stemmer

The Japanese stemmer follows the general stemming rules, described in Stemming. In brief, the major word classes, also known as the open classes, stem to their base or dictionary forms. This is shown in the table below.

Category	Baseform
Noun	Source form
Verb	Non-past-tense form
Adjective	Non-past-tense form
Adverb	Source form

Classifiers and derivational suffixes are not removed from nouns in the stemmer. For example,

Part of Speech	Word	Stem
`Num + Cl`	二冊	二冊
`Nn + Adj_suffix`	高さ	高い
`Verb + Nn_suffix`	読み方	読み方
`Nn + Pl_suffix`	学生たち	学生たち
`Nn + Hon_suffix`	佐藤様	佐藤様

Closed class words like pronouns, demonstratives, letters and numbers are stemmed to their base form, and any existing case markers are dropped.

Japanese verb and adjective words can be inflected for tense, aspect, modality, politeness, and so on. The stemmer returns verbs and adjectives without any inflectional endings. This is commonly called the dictionary form. For example, the following inflected verb forms all stem to 食べる ("eat").

Word	Stem
食べた	食べる
食べさせる	食べる
食べます	食べる
食べない	食べる

5.14.1.3.2 Expanded Stemmer

The expanded Japanese language module provides more fine-grained segmentation and stemming results than the standard module. Its output is designed for optimized text indexing and search systems. The expanded module output differs from the standard stemmer in that classifiers, numerals, prefixes and suffixes are separated from their head words, and compound analysis is performed.

Examples are shown below.

Classifiers are separated from numerals:

Text	Output
1996年	1996
	年
30 分	30
	分

Prefixes are separated from their head words:

Text	Output
お部屋	お
	部屋
副作用	副
	作用

Suffixes are separated from their head words:

Text	Output
全国的	全国
	的
須田さん	須田
	さん
ニューヨーク州	ニューヨーク
	州

Compounds are broken into their separate components:

Text	Output
朝日新聞社	朝日
	新聞
	社

Text	Output
日本電信電話	日本
	電信
	電話
サウンドマスター	サウンド
	マスター

The expanded variant supports all the same operations as the standard Japanese module. However, its fine-grained output provides less contextual information for each term, and this ambiguity can compromise the accuracy of the tagging operations. For these operations, we recommend using the standard Japanese module. The expanded variant is recommended for stemming purposes only.

5.14.1.4 Part-of-Speech Tagging in Japanese

The following table shows the Japanese tag set. The tag names are accompanied by a brief description and one or more examples.

Umbrella Tag	Complete Tag	Description	Examples
Adj	Adj	Adjective	赤い、大きい
	Adj-D	Adjective denoting modality	（て）ほしい、（て）よい
Adnom	Adnom	Prenominal nominal	この、そんな
Adv	Adv	Adverb	ゆっくり、じっと
Aux	Aux	Auxiliary verb	だ、です、ない
Case	Case	Case marker	が、を、さえ
Conj	Conj	Conjunction	そして、しかし
Interj	Interj	Interjection	さあ、えっ

Umbrella Tag	Complete Tag	Description	Examples
Nn	Nn	Noun	先生、分析、ファイル
	Nn-Adv	Noun commonly used adverbially	今日、午後、１月
	Nn-Ascii	ASCII character, sequence or word	computer
	Nn-D	Formal noun; non-contentful noun	こと、の、もの, よう
	Nn-Pron	Pronoun	あなた、私、ここ
	Nn-Prop	Proper noun	山田、富士山
Num	Num	Numeric nominal	２０００年、９５％
Pre	Pre	Nominal prefix	お（水）、高（品質）
Punct	Punct	Punctuation	”：＃＠
	Punct-Close	Closing punctuation	）、}、」
	Punct-Comma	Comma	、
	Punct-Open	Opening punctuation	（、{、「
	Punct-Sent	Sentence-ending punctuation	。？
Suf	Suf	Suffix	さん
Verb	Verb	Verb	読む、理解する
	Verb-D	Verb denoting modality and aspect	はじめる、できる、（て）いる

5.14.2 Extraction

This section describes the extraction-specific information for Japanese.

5.14.2.1 Japanese Subtypes

Japanese supports subtypes in the types: ORGANIZATION, REGION, and URI.

5.14.2.2 Predefined Entity Types

This section describes the predefined entity types supported by the Japanese language module and examples of each. Click on any link to jump to that section: ADDRESS1, CONTINENT, COUNTRY, CURRENCY, DATE, DAY, GEO_AREA, GEO_FEATURE,HOLIDAY, LANGUAGE, LOCALITY, MEASURE, MONTH, NAME_DESIGNATOR, NOUN_GROUP, ORGANIZATION, PEOPLE, PERCENT, PERSON, PHONE, PRODUCT, PROP_MISC, REGION, TIME, TIME_PERIOD, TITLE, URI, and YEAR.

5.14.2.2.1 ADDRESS1

Postal addresses. At least two pieces of the address information need to be adjacent to get ADDRESS1 tag:

- 540-8505大阪府大阪市中央区城見2丁目2番53号
- 329-1194栃木県宇都宮市下岡本町2145番13号

5.14.2.2.2 CONTINENT

Any of the continents, for example:

- アジア
- 欧州

5.14.2.2.3 COUNTRY

Name of the countries and the names of geo-political entities for which the conventional labels do not apply. For example:

- 日本
- フランス
- 台湾

5.14.2.2.4 CURRENCY

Expressions denoting amounts of money, for example:

- ４７円
- 65ドル

5.14.2.2.5 DATE

Dates are minimally composed of a number and month name:

- 7月26日
- 2012/12/20

5.14.2.2.6 DAY

Names of the days of the week:

- 火曜日
- にちようび

5.14.2.2.7 GEO_AREA

A geographical area that captures a significant land mass, such as a group of countries. For example:

- 東南アジア
- スラヴ諸国
- 五大湖地方

5.14.2.2.8 GEO_FEATURE

The names of locations that are geologically or ecosystemically designed, non-artificial locations or bodies of water. For example:

- 富士山
- 国後島
- インド洋

5.14.2.2.9 HOLIDAY

Holidays and special days:

- 天皇誕生日
- こどもの日
- クリスマス

5.14.2.2.10 LANGUAGE

Name of a language:

- 日本語
- ロシア語

5.14.2.2.11 LOCALITY

Name of a city:

- 横浜
- 上海市

5.14.2.2.12 MEASURE

Measure expressions:

- 60km
- １６ミリ

5.14.2.2.13 MONTH

Names of the months of the year:

- 7月
- 七月

5.14.2.2.14 NAME_DESIGNATOR

Name designator 気付 that appears after a person's name:

- スミス様気付
- 雅子さま気付

5.14.2.2.15 NOUN_GROUP

Noun groups can be simple or compound nouns with modifying adjectives:

- 乳製品事業
- 企業理念

5.14.2.2.16 ORGANIZATION

Government, legal, and service agencies, including non-profit organizations, fine arts groups, and other associations and institutions, extracted as one of the following subtypes:

- COMMERCIAL–The name of commercial organizations, such as major companies or corporations, for example:
  - マイクロソフト
  - 株式会社豊田自動織機

  Note:
  Any extracted ORGANIZATION/COMMERCIAL entities can be parsed and standardized using the Data Quality Data Cleanse transform by mapping them to one of the FIRM input fields.

- EDUCATIONAL–The names of institutions focused primarily in education, for example:
  - 関東学院
  - マギル大学

- OTHER–Any organization that does not fit into a more specific subtype, for example:
  - 自民党
  - ニューヨーク・ ヤンキース

5.14.2.2.17 PEOPLE

Names referring to nationalities, for example:

- 日本人
- アメリカ人

5.14.2.2.18 PERCENT

Percent expressions:

- 25％
- 10パーセント

5.14.2.2.19 PERSON

Variations of person names, for example:

- オバマ
- 菅直人

5.14.2.2.20 PHONE

Phone numbers based on the Japanese format, for example:

- 電話： 0566-52-2120
- （電）０８６・ ２２４・ ４３２０

5.14.2.2.21 PRODUCT

Name of a product, for example:

- 一太郎
- アイパッ

5.14.2.2.22 PROP-MISC

Some other proper noun phrases that do not belong to one of the entity types specified by the other entities, for example:

- ナスダック総合指数
- 万国博覧会

5.14.2.2.23 REGION

Different regions are extracted as the following subtype:

- MAJOR—The major administrative divisions of countries, such as the prefectures of Japan and the states of the United States. For example:
  - 東京都
  - 岩手県
  - カリフォルニア州

5.14.2.2.24 TIME

Clock times and time expressions, for example:

- 18時49分
- 正午

5.14.2.2.25 TIME_PERIOD

Measures of time duration, for example:

- 二時間
- ４か月

5.14.2.2.26 TITLE

Names of important positions in government, business, and other organizations: such as:

- 大統領
- 名誉教授

5.14.2.2.27 URI

An address on the internet, extracted as one of the following subtypes:

- EMAIL—Email addresses, for example:
  - john.doe@hotmail.com
  - somebody@sap.com
- IP—IP addresses, for example:
  - 169.135.52.80
- URL—Internet address, for example:
  - http://www2.toyota.co.jp/
  - www.sap.com

5.14.2.2.28 YEAR

A year identifier and expressions based on years, for example:

- 2012年
- 平成23年

5.15 Korean Language Reference

This chapter describes the behavior of the Korean language module.

5.15.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Korean texts, including word segmentation and stemming.

5.15.1.1 Character Encodings for Korean

- euc_kr
- utf_8, utf_16, ucs_4

5.15.1.2 Word Segmentation in Korean

The Korean segmenter uses the same underlying algorithm as the segmenters for the white space languages. In brief, punctuation characters are treated as syntactic unit delimiters, as are the white space and tab characters. The Korean segmenter has the following language-specific behavior.

Korean words can contain several sorts of dependent morphemes, such as case markers and inflectional endings. The dependent morphemes do not become separate words.

For example, noun 사람들은 "person-PL-TOP" is segmented as one word even though it consists of three morphemes, the noun 사람 , the plural marker 들, and the topic marker 은 . Similarly, in 가셨습니다 ("a (respectable person) has gone"), the subject honorific 시 , the past tense suffix 었 , which

contract together as 셨 the addressee honorific suffix 습, the indicative suffix 니, and the declarative suffix 다 occur in that order after the head verb stem 가 ("go").

Text	Segmented
그	그
사람들은	사람들은
못됐다	못됐다
.	.

Text	Segmented
선생님께서는	선생님께서는
벌써	벌써
서울로	서울로
가셨습니다	가셨습니다
.	.

Segmenters for European languages recognize multiword units as a single unit (for example, "to and fro" in English). The Korean segmenter gives the same treatment to phrases like "이랬다" and "저랬다".

Related Topics

- Word Segmentation

5.15.1.3 Stemming in Korean

The Korean stemmer follows the general stemming rules, outlined in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms, or citation forms. This is shown in the table below. Closed class words like pronouns are also be stemmed in Korean, but they are not shown here. Uninflected forms like determiners stem to themselves.

The Korean system supports two stemmers: Korean standard stemmer and Korean expanded stemmer. The primary difference is that the expanded stemmer breaks compound nouns while the standard stemmer does not. If stem guessing is important to your application, it is recommend that you run the expanded and standard stemmers such that the expanded stemmer is consulted first.

5.15.1.3.1 Standard Stemmer

Category	Baseform
Noun/Pronoun	Base form (without case marking)
Verb	Declarative form
Adjective	Declarative form
Adverb	Source form

A Korean noun word may include a postposition (particle) indicating case marking. The stemmer returns the normalized noun (such as the uninflected noun head or content word) minus any case marking.

Nouns, pronouns, proper nouns and numerals all stem to the base form, without any case markers. For example, the following nouns are all stemmed to 학생:

Word	Stem
학생이	학생
학생을	학생
학생까지	학생
학생한테서	학생
학생하고	학생

Word	Stem
학생까지만	학생

Verbs and adjectives stem to the dictionary form, without any inflectional suffixes. The following inflected verb forms are all stemmed to 먹다 ("eat"):

Word	Stem
먹었다	먹다
먹었겠다	먹다

The sentence 학생이 케이크 마지막 조각을 먹었다 . ("The student has eaten the last piece of cake.") is stemmed as follows:

Word	Stem
학생이	학생
케이크	케이크
마지막	마지막
조각을	조각
먹었다	먹다

5.15.1.3.2 Korean Compound Analysis

Compound stemming in Korean takes place in the expanded inflectional stemmer module, using `korean-expanded.stemmer`. While this dictionary has the same name as the expanded inflectional stemmers in the European languages, it performs compound stemming in Korean. The Korean module handles two-part compounds of two types: Noun-Verb and Noun-Noun.

Note:
Due to the complex internal structure of compounds in Korean, the expanded inflectional stemming operation may take relatively more time than other operations.

The sample output below uses the vertical bar (|) to delimite terms or stems. Compounds are always broken up.

Noun-Verb Compounds

A noun combines with an intransitive verb to create a compound verb. The noun is uninflected, and the verb can be inflected and will be stemmed to its base form.

Example	Output
계획이다	계획 | 이다
기술적인	기술적 | 이다
시간두고	시간 | 두다
해결되다	해결 | 되다

Noun-Noun Compounds

A noun combines with a noun to create another compound noun. The first noun is uninflected, and the second noun can be inflected and will be stemmed to its base form.

Example	Output
연기상을	연기 | 상
현대문제가	현대 | 문제
거래소시장에서	거래소 | 시장

5.15.1.4 Part-of-Speech Tagging in Korean

The following table shows the Korean tag set available for using in defining custom entities. The tag names are accompanied by a brief description and one or more examples.

Umbrella Tag	Complete Tag	Description	Examples
`Adv`	`Adv`	Adverb	만일, 그러면
`Case`	`Case`	Case Marker	을, 를
`Det`	`Det`	Determiner	이, 그
`Interj`	`Interj`	Interjection	아이고, 어머

Umbrella Tag	Complete Tag	Description	Examples
Nn	Nn	Noun	책, 코끼리
	Nn-Ascii	ASCII characters, Noun	copyright, Web
	Nn-Case	Noun Case Marked	호박은
	Nn-Case-Acc	Noun Case Marked Accusative	조각을
	Nn-Case-Conj	Conjoined Case Marked Noun	고양이와는, 토끼와는
	Nn-Case-Conj-Pl	Conjoined Case Marked Plural Noun	교인들하고는
	Nn-Case-Disj	Disjunctive Case Marked Noun	여기까지나
	Nn-Case-Disj-Pl	Disjunctive Case Marked Plural Noun	박사들까지나
	Nn-Case-Gen	Noun Case Marked Genitive	인간의
	Nn-Case-Pl	Noun Case Marked-Plural	선생님들은, 군인들은
	Nn-Case-Pl-Acc	Noun Case Marked-Plural Accusative	학생들을
	Nn-Case-Pl-Gen	Noun Case Marked-Plural Genitive	교수들의
	Nn-Conj	Conjoined Noun	강아지와, 사자와
	Nn-Conj-Pl	Conjoined Plural Noun	교인들하고, 친구들하고
	Nn-Disj	Disjunctive Noun	짐승이나, 과학자나
	Nn-Disj-Pl	Disjunctive Plural Noun	약사들이나, 화가들이나
	Nn-Pl	Noun-Plural	사람들, 박사들
Num	Num	Numeric Expression	30분, 삼십분

Umbrella Tag	Complete Tag	Description	Examples
Pron	Pron	Pronoun	나, 너
	Pron-Case	Case Marked Pronoun	나는
	Pron-Case-Acc	Pronoun Case Marked Accusative	너를
	Pron-Case-Conj	Pronoun Case Marked-Conjunctive	자네하고는
	Pron-Case-Conj-Pl	Pronoun Case Marked-Conjunctive Plural	우리들하고는
	Pron-Case-Disj	Pronoun Case Marked-Disjunctive	자기나만은
	Pron-Case-Disj-Pl	Pronoun Case Marked-Disjunctive Plural	저이들까지나
	Pron-Case-Gen	Pronoun Case Marked Genitive	나의
	Pron-Case-Pl	Pronoun Case Marked Plural	우리들은
	Pron-Case-Pl-Acc	Pronoun Case Marked Plural Accusative	우리들을
	Pron-Case-Pl-Gen	Pronoun Case Marked Plural Genitive	우리들의
	Pron-Conj	Conjoined Pronoun	자네하고
	Pron-Conj-Pl	Conjoined Plural Pronoun	우리들하고
	Pron-Disj	Disjunctive Pronoun	자기나
	Pron-Disj-Pl	Disjunctive Plural Pronoun	우리들이나
	Pron-Pl	Plural Pronoun	우리들, 그들

Umbrella Tag	Complete Tag	Description	Examples
Prop	Prop	Proper Name	삼성전자, 서울대학교
	Prop-Case	Case Marked Proper Name	현대건설은
	Prop-Case-Acc	Proper Name Case Marked Accusative	고려대학교를
	Prop-Case-Conj	Proper Name Case Marked Conjunctive	나이키와는
	Prop-Case-Disj	Proper Name Case Marked Disjunctive	소니에게나
	Prop-Case-Gen	Proper Name Case Marked Genitive	한국은행의
	Prop-Conj	Conjoined Proper Name	동국제강과
	Prop-Disj	Disjunctive Proper Name	UBS나
Punct	Punct	Punctuation	;, '
	Punct-Comma	Punctuation-Comma	,
	Punct-Sent	Punctuation-Sentence	.
V	V-Fut	Future Tense Verb	판매하겠다, 시작하겠다
	V-Past	Past Tense Verb	출발했다, 몰랐었다
	V-PreMod	Pre-modifying Verb	좋은
	V-Pres	Present Tense Verb	상회하다, 번거롭다

5.15.1.4.1 Unfound Words

Words not found in the tagger dictionary are passed to the Korean guesser where they are assigned a tag based on a set of rules about Korean morphology. For instance, it relies on noun suffixes to determine a `Nn-*` tag. Foreign words, such as English words, are tagged `Nn-Ascii`.

5.15.2 Extraction

This section describes the extraction-specific information for Korean.

5.15.2.1 Korean Subtypes

Korean supports subtypes in the types `ORGANIZATION`, `REGION`, `SOCIAL_MEDIA`, and `URI`.

Related Topics

- Subtypes

5.15.2.2 Predefined Entity Types

This section describes the predefined entity types supported by the Korean language module and examples of each. Click each link to jump to that subsection: COUNTRY, FACILITY, GEO_AREA, GEO_FEATURE, LOCALITY, ORGANIZATION, PERSON, PHONE, REGION, SOCIAL_MEDIA, TITLE, and URI

5.15.2.2.1 COUNTRY

Names of countries:

- 중국이
- 일본은
- 러시아

5.15.2.2.2 FACILITY

Man-made structures:

- 주공아파트
- 롯데호텔에서
- 성수대교를

5.15.2.2.3 GEO_AREA

A geographical area larger than a city that captures a significant land mass, such as a continent or a group of countries:

- 스페인 서쪽

- 이라크 남동부
- 동아시아

5.15.2.2.4 GEO_FEATURE

The names of locations such as borders, astronomical locations, bodies of water, or locations that are geologically or ecosystemically designed:

- 지구로부터
- 달천강
- 남극해

5.15.2.2.5 LOCALITY

Names of cities:

- 서울
- 로마나
- 모스크바의

5.15.2.2.6 ORGANIZATION

Government, legal, and service agencies, including non-profit organizations, fine arts groups, and other associations and institutions, extracted as one of the following subtypes:

- `COMMERCIAL`–The name of commercial organizations, such as major companies or corporations. For example:
  - 동국제강
  - 삼성전자는
  - 현대자동차가

  Note:
  Any extracted `ORGANIZATION/COMMERCIAL` entities can be parsed and standardized using the Data Quality Data Cleanse transform by mapping them to one of the `FIRM` input fields.
- `OTHER`–Any organization that does not fit into a more specific subtype.
  - IAEA가
  - 경원대에서
  - 청와대가

5.15.2.2.7 PERSON

Variations of names:

- 노무현
- 김대중을
- 한승주에게

5.15.2.2.8 PHONE

Phone numbers:

- 02-3321-2345
- (011)222-3456
- 1-800-223-4567

5.15.2.2.9 REGION

Different regions are extracted as one of the following subtypes:

- `MAJOR`– Names of states and provinces:
  - 제주도
  - 경기도와
  - 하와이
- `MINOR`– Names of counties, prefectures, districts, or analogous geographical divisions or governmental units:
  - 금산군
  - 종로구에서
  - 관악구

5.15.2.2.10 SOCIAL_MEDIA

Entity type for extracting entities from social media feeds. The handles (also known as `ID`) and topics are extracted as one of the following subtypes:

Note:
The SOCIAL_MEDIA entity type supports only Twitter feeds.

- `ID_TWITTER`–Twitter handles or `ID`s starting with "`@`", for example:
  - @HyunheeJeon
  - @SangSangYi
  - @갤탭inkorea10
  - @SecretGarden_KD
- `TOPIC_TWITTER`–Twitter topics starting with "`#`", for example:

- #서양이동양에게삶을묻다_
- #갤탭10
- #JBBANK
- #JP가10

5.15.2.2.11 TITLE

Names of important positions in government, business, and other organizations:

- 장관을
- 대통령이
- 상임위원장

5.15.2.2.12 URI

An address on the internet, extracted as one of the following subtypes:

- `EMAIL`–Email addresses, for example:
  - abc_333@sun.com
  - smlee@yna.co.kr
  - Jesus Melendrez/Corp/Enron@CMP
- `URL`–Internet addresses, for example:
  - www.cyworld.com/common
  - http://www.cnn.com/2007/US/law/07/17/couey.hearing/index.html
  - http://kr.news.yahoo.com/service/news/shellsection.htm?linkid

5.16 Norwegian: Bokmål Language Reference

This chapter describes the behavior of the Bokmål language module.

5.16.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Bokmål texts, including word segmentation and stemming.

5.16.1.1 Character Encodings for Bokmål

- iso_8859_1
- cp_1252
- utf_8, utf_16, ucs_4

5.16.1.2 Word Segmentation in Bokmål

The Bokmål segmenter follows all of the general segmentation rules in the white space languages. The Bokmål segmenter has the following language-specific behavior.

The Bokmål segmenter does not split plurals and possessives spelled with **s** or **'s**. Hyphens are not separated from compound parts written with a hyphen. Periods are not separated from ordinal digit expressions.

Text	Segmented
Eriks	Eriks
32.	32.
lonns- og inntektsutviklingen	lonns-
	og
	inntektsutviklingen

Related Topics

- Word Segmentation

5.16.1.3 Stemming in Bokmål

This section describes the standard stemmer and the expanded inflectional stemmer used for stemming in Bokmål.

5.16.1.3.1 Standard Stemmer

The Bokmål stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Baseform	Examples
Noun	Indefinite singular	dammer -> dam; bondens -> bonde
Verb	Infinitive	ventet -> vente; sendes -> sende
Adjective	Base form	laveste -> lav; kalde -> kald
Adverb	Base form or source form	nærest -> nær; imens -> imens

5.16.1.3.2 Expanded Inflectional Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text (such as email, online documents, or queries). See Expanded Inflectional Stemming for general behavior. . The specifics for Bokmål follow.

Case Variants

The expanded version acceps lower case letters in addition to capital letters for words that are usually capitalized.

Example	Output
Erik	Erik

Example	Output
erik	Erik

Typewriter Forms of Accented Letters

The expanded version accepts typewriter conventions for accented letters. That is, **å** is recognized when written as **aa**, **æ** when written as **ae**, and **ø** when written as **oe**.

Example	Output
blaa	blå
blå	blå

Deaccented Characters

The expanded version also allows deaccented characters in place of accented ones.

Example	Output
bla	blå
blå	blå

Hyphenation

To aid software that handles line-breaking hyphens by deleting them and concatenating the two parts of the broken word, hyphens in non-numeric expressions are optional in the expanded version, so that words that are truly hyphenated will still be recognized.

Example	Output
Nord-Vestlandet	Nord-Vestlandet
NordVestlandet	Nord-Vestlandet

5.16.1.4 Part-of-Speech Tagging in Bokmål

The following table shows the Norwegian Bokmål tag set available for defining custom entities. The tag names are accompanied by a brief description and one or more examples. If the example consists of more than one word, the word exemplifying the current tag is in **bold**. The tag set makes no distinction for gender.

Umbrella Tag	Complete Tag	Description	Examples
`Abbr`	`Abbr`	Abbreviation	adr., ibid.

Umbrella Tag	Complete Tag	Description	Examples
Adj	Adj-Comp	Comparative adjective	mindre, eldre
	Adj-Comp-Gen	Genitive comparative adjective	eldres
	Adj-Def-Pl	Definite and/or plural adjective	gretne, mørke, kalde
	Adj-Def-Pl-Gen	Genitive definite and/or plural adjective	fremmedes, uvørnes
	Adj-Def-Sg	Definite singular adjective	lille
	Adj-Def-Sg-Gen	Genitive definite singular adjective	lilles
	Adj-Indef-Sg	Indefinite singular adjective	skriftlig
	Adj-Indef-Sg-Gen	Genitive indefinite singular adjective	skriftligs
	Adj-PaPart-Def-Pl	Definite and/or plural adjectival past participle	besøkte, befestede, dempede
	Adj-PaPart-Def-Pl-Gen	Genitive, definite and/or plural, adjectival past participle	besøktes, befestedes, dempedes
	Adj-PaPart-Indef-Sg	Adjectival past participle, indefinite singular	for **kryptert** kommunikasjon
	Adj-Pl	Plural adjective	**små** problemer
	Adj-Pl-Gen	Genitive plural adjective	smås
	Adj-PrPart	Adjectival present participle	begynnende, stirrende
	Adj-PrPart-Gen	Genitive adjectival present participle	reisendes, lekendes, gråtendes
	Adj-Sup	Superlative adjective	best, raskest, størst
	Adj-Sup-Def	Definite superlative adjective	fineste, innerste, viktigste

Umbrella Tag	Complete Tag	Description	Examples
	Adj-Sup-Def-Gen	Genitive definite superlative adjective	finestes, innerstes, viktigstes
Adv	Adv	Adverb	ikke, oppe, heller
	Adv-Comp	Comparative adverb	lenger
	Adv-Sup	Superlative adverb	helst
Aux	Aux/V-Impv	Imperative auxiliary or main verb	bli
	Aux/V-Inf	Infinitive auxiliary or main verb	vaere
	Aux/V-Inf-SForm	S-Form of infinitive auxiliary or main verb	has
	Aux/V-PaPart	Past participle of auxiliary or main verb	hatt, vaert, blitt, fått
	Aux/V-Past	Past tense auxiliary or main verb	hadde, var, ble
	Aux/V-Pres	Present tense auxiliary or main verb	har, er, blir, får
	Aux/V-PrPart	Present participle of auxiliary or main verb	havende, blivende
	Aux-Inf	Infinitive auxiliary verb	kunne, måtte
	Aux-Inf-SForm	S-Form of infinitive auxiliary verb	kunnes, måttes
	Aux-PaPart	Past participle of auxiliary verb	kunnet, måttet
	Aux-Past	Past tense auxiliary verb	kunne, måtte
	Aux-Pres	Present tense auxiliary verb	kan, vil
	Aux-Pres-SForm	S-Form of present tense auxiliary verb	villes, skulles
	Aux-PrPart	Present participle of auxiliary verb	villende, skullende

Umbrella Tag	Complete Tag	Description	Examples
Cmpd	Cmpd-Part	Left compound part	**kontor** - og forretnings-bygg
Conj	Conj	Subordinating or relativizing conjunction	som, mens
	Conj-Coord	Coordinating conjunction	og, eller
Det	Det-Art-Def-Pl	Definite plural determiner	disse
	Det-Art-Def-Sg	Definite singular determiner	på **denne** grunn
	Det-Art-Indef	Indefinite singular determiner	en, et
	Det/Pron-Comp	Determiner or pronoun, comparative	mer
	Det/Pron-Pl	Determiner or pronoun, plural	noen, hvilke, alle, andre
	Det/Pron-Pl-Gen	Determiner or pronoun, plural genitive	noens, alles, andres
	Det/Pron-Sg	Determiner or pronoun, singular	hver, litt, alt
	Det/Pron-Sg-Gen	Determiner or pronoun, singular genitive	enhvers, annens
	Det/Pron-Sup	Determiner or pronoun, superlative	mest
	Det/Pron-Sup-Def	Determiner or pronoun, definite superlative	meste
Interj	Interj	Interjection	ja, herregud

Umbrella Tag	Complete Tag	Description	Examples
Nn	Nn-Def-Pl	Definite plural noun	dørene, armene
	Nn-Def-Pl-Gen	Genitive definite plural noun	salongenes, kollegenes
	Nn-Def-Sg	Definite singular noun	flyet, klokken
	Nn-Def-Sg-Gen	Genitive definite singular noun	selskapets, spisestuens
	Nn-Indef-Pl	Indefinite plural noun	plasser, mapper
	Nn-Indef-Pl-Gen	Genitive indefinite plural noun	tiders, menneskers
	Nn-Indef-Sg	Indefinite singular noun	stol, stripe
	Nn-Indef-Sg-Gen	Genitive indefinite singular noun	topps
	Nn-Indef-SP	Indefinite singular or plural noun	lys, skritt
	Nn-Indef-SP-Gen	Genitive indefinite singular or plural noun	slags, lands, års
	Nn-Letter	Lowercase and uppercase letters	b, N
	Nn-Net	URL and e-mail address	www.inxight.com in fo@inxight.com
Num	Num	Cardinal numeric expression or plural cardinal number (spelled out)	-294, 4,6%, xii, 1.100to, tre, fire
	Num-Def-Sg	The number "one", definite singular (spelled out)	ene
	Num-Indef-Sg	The number "one", indefinite ssingular (spelled out)	en, ett
Ord	Ord	Ordinal number (in digits or spelled out)	7., første
Part	Part-Inf	Infinitival particle	**å** beskrive

Umbrella Tag	Complete Tag	Description	Examples
Prep	Prep	Preposition	med, ut
	Prep-av	Preposition **av**	av
	Prep-for	Preposition **for**	for
	Prep-fra	Preposition **fra**	fra
	Prep-i	Preposition **i**	i
	Prep-paa	Preposition **på**	**på** bakgrunn
	Prep-ved	Preposition **ved**	ved
Pron	Pron-Acc	Accusative pronoun	ham, henne
	Pron-Nom	Nominative pronoun	han, hun
	Pron-Poss-Pl	Possessive pronoun with plural agreement	sine
	Pron-Poss-Sg	Possessive pronoun with singular agreement	sin
Prop	Prop	Proper name	Oslo, Arne
	Prop-Gen	Genitive proper name	Akers
Punct	Punct	Miscellaneous punctuation	- [
	Punct-Comma	Comma	,
	Punct-Quote	Quotation marks	" ' ' " "
	Punct-Sent	Sentence boundary punctuation	. ... ? : ; !

Umbrella Tag	Complete Tag	Description	Examples
V	V-Impv	Imperative verb	se, ta
	V-Inf	Infinitive verb	komme, gjøre
	V-Inf-SForm	S-Form of infinitive verb	kan **belastes**
	V-PaPart	Past participle verb	reist, utpekt, stanset
	V-PaPart-SForm	S-Form of past participle verb	trivdes
	V-Past	Past tense verb	sa, vokste
	V-Past-SForm	S-Form of past tense verb	levdes, mistrivdes
	V-Pres	Present tense verb	vet, gir
	V-Pres-SForm	S-Form of present tense verb	flyttes, møtes, finnes, synes
	V-PrPart	Present participle verb	være **avtakende**

5.16.1.5 Grouping in Bokmål

A Bokmål simple noun phrase may consist of one or more nouns or proper nouns, as in:

- Arne Huuse

Bokmål noun groups can include various modifiers, such as adjectives, possessives, and indefinite nouns, for example:

- nordisk rett
- utvalgets sekretær
- statsadvokat Ketil Haukaas

Noun phrases also include compound parts or can be coordinated with **og** or **eller**, such as:

- person- og rettsvern
- politi og påtalemyndighet
- subsumsjon eller straffutmåling

A simple noun phrase may also combine with prepositional phrases that start with the prepositions **av** and **fra**, for example:

- formidling av informasjon
- instrukser fra riksadvokaten

Noun phrases also include the prepositions **ved**, **i**, and **på**, when they're followed by a proper noun:

- kontrollen ved Norsk Tipping AS
- kasino i Finland
- organisasjon på Østlandet

Noun phrases include the preposition for when it follows proper nouns, as in:

- Internett for privatpersoner

5.16.2 Extraction

This section describes the extraction-specific information for Bokmål.

5.16.2.1 Predefined Entity Types

This section describes the predefined entity types supported by the Bokmål language module and examples of each.

5.16.2.1.1 NOUN_GROUP

A Bokmål simple noun phrase may consist of one or more nouns or proper nouns, as in:

- Arne Huuse

Bokmål noun groups can include various modifiers, such as adjectives, possessives, and indefinite nouns, for example:

- nordisk rett
- utvalgets sekretær
- statsadvokat Ketil Haukaas

Noun phrases also include compound parts or can be coordinated with **og** or **eller**, such as:

- person- og rettsvern

- politi og påtalemyndighet
- subsumsjon eller straffutmåling

5.17 Norwegian: Nynorsk Language Reference

This chapter describes the behavior of the Nynorsk language module.

5.17.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Nynorsk texts, including word segmentation and stemming.

5.17.1.1 Character Encodings for Nynorsk

- iso_8859_1
- cp_1252
- utf_8, utf_16, ucs_4

5.17.1.2 Word Segmentation in Nynorsk

The Nynorsk segmenter follows all of the general segmentation rules in the white space languages. The Nynorsk segmenter has the following language-specific behavior.

The Nynorsk segmenter does not split plurals and possessives spelled with **s** or **'s**. Hyphens are not separated from compound parts written with a hyphen. Periods are not separated from ordinal digit expressions.

Text	Segmented
Eriks	Eriks
32.	32.
lonns- og inntektsutviklinga	lonns-
	og
	inntektsutviklinga

Related Topics

- Word Segmentation
- White Space Languages

5.17.1.3 Stemming in Nynorsk

This section describes the standard stemmer and the expanded inflectional stemmer used for stemming in Nynorsk.

5.17.1.3.1 Standard Stemmer

The Nynorsk stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Baseform	Examples
Noun	Indefinite singular	bilen -> bil; jenta -> jente
Verb	Infinitive	leikte -> leike; speil -> speile
Adjective	Base form	høgare -> høg; blått -> blå

Category	Baseform	Examples
Adverb	Base form or source form	svintare -> svint; imedan -> imedan

5.17.1.3.2 Expanded Inflectional Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text (such as email, online documents, or queries). See Expanded Inflectional Stemming for general behavior. The specifics for Nynorsk follow.

Case Variants

The expanded version accepts lower case letters in addition to capital letters for words that are usually capitalized.

Example	Output
Erik	Erik
erik	Erik

Typewriter Forms of Accented Letters

The expanded version accepts typewriter conventions for accented letters. That is, **å** is recognized when written as **aa**, **æ** when written as **ae**, and **ø** when written as **oe**.

Example	Output
blaa	blå
blå	blå

Deaccented Characters

The expanded version also allows deaccented characters in place of accented ones.

Example	Output
bla	blå
blå	blå

Hyphenation

To aid software that handles line-breaking hyphens by deleting them and concatenating the two parts of the broken word, hyphens in non-numeric expressions are optional in the expanded version, so that words that are truly hyphenated will still be recognized.

Example	Output
NATO-land	NATO-land
NATOland	NATO-land

5.17.1.4 Part-of-Speech Tagging in Nynorsk

The following table shows the Norwegian Nynorsk tag set available for defining custom entities. The tag names are accompanied by a brief description and one or more examples. If the example consists of more than one word, the word exemplifying the current tag is in **bold**. The tag set makes no distinction for gender.

Umbrella Tag	Complete Tag	Description	Examples
`Abbr`	`Abbr`	Abbreviation	red.

Umbrella Tag	Complete Tag	Description	Examples
Adj	Adj-Comp	Comparative adjective	mindre, eldre
	Adj-Def-Pl	Definite and/or plural adjective	gretne, mørke, kalde
	Adj-Def-Pl-Gen	Genitive definite plural adjective	framandes
	Adj-Def-Sg	Definite singular adjective	lisle, vesle
	Adj-Indef-Sg	Indefinite singular adjective	norsk
	Adj-Indef-Sg-Gen	Genitive indefinite singular adjective	nærliggjandes
	Adj-PaPart-Def-Pl	Definite and/or plural adjective derived from past participle	kjende
	Adj-PaPart-Indef-Sg	Indefinite singular adjective derived from past participle	reist
	Adj-Pl	Plural adjective	ørsmåe
	Adj-PrPart	Adjective derived from present participle	begynnande, stirande
	Adj-PrPart-Gen	Genitive adjective derived from present participle	reisandes, leikandes
	Adj-Sup	Superlative adjective	best, raskast, størst
	Adj-Sup-Def	Definite superlative adjective	finaste, inste, viktigaste
Adv	Adv	Adverb	ikkje, no
	Adv-Comp	Comparative adverb	lenger, heller
	Adv-Sup	Superlative adverb	verst

Umbrella Tag	Complete Tag	Description	Examples
Aux/V	Aux/V-Impv	Imperative auxiliary or main verb	ver
	Aux/V-Inf	Infinitive auxiliary or main verb	bli
	Aux/V-Inf-SForm	S-Form of infinitive auxiliary or main verb	havast, fåast
	Aux/V-PaPart	Past participle of auxiliary or main verb	hatt, vore, blitt, fått
	Aux/V-Past	Past tense auxiliary or main verb	hadde, var, blei, fekk
	Aux/V-Pres	Present tense auxiliary or main verb	har, er, blir, får
	Aux/V-PrPart	Present participle of auxiliary or main verb	havande, blivande, verande
Aux	Aux-Inf	Infinitive auxiliary verb	kunne, måtte
	Aux-PaPart	Past participle of auxiliary verb	vilja, måtta
	Aux-Past	Past tense auxiliary verb	kunne, måtte
	Aux-Pres	Present tense auxiliary verb	kan, vil
	Aux-PrPart	Present participle of auxiliary verb	viljande, kunnande
Cmpd	Cmpd-Part	Left compound part	**kontor** - og forretnings-bygg
Conj	Conj	Subordinating or relativizing conjunction	som, mens
	Conj-Coord	Coordinating conjunction	og, eller

Umbrella Tag	Complete Tag	Description	Examples
Det/Pron	Det/Pron-Comp	Comparative determiner or pronoun	meir
	Det/Pron-Pl	Plural determiner or pronoun	alle
	Det/Pron-Sg	Singular determiner or pronoun	nokon
	Det/Pron-Sup	Superlative determiner or pronoun	mest
	Det/Pron-Sup-Def	Definite superlative determiner or pronoun	meste
Det	Det-Art-Def-Pl	Definite plural determiner (article or demonstrative pronoun)	dei, desse
	Det-Art-Def-Sg	Definite singular determiner (article or demonstrative pronoun)	**denne** artikkelen
	Det-Art-Indef	Indefinite singular determiner	eit
Interj	Interj	Interjection	hei, tjo

Umbrella Tag	Complete Tag	Description	Examples
Nn	Nn-Def-Pl	Definite plural noun	dørene, armane
	Nn-Def-Pl-Gen	Genitive definite plural noun	salonganes, kollegaenes
	Nn-Def-Sg	Definite singular noun	flyet, klokka
	Nn-Def-Sg-Gen	Genitive definite singular noun	selskapets, stovas
	Nn-Indef-Pl	Indefinite plural noun	plassar, mapper
	Nn-Indef-Pl-Gen	Genitive indefinite plural noun	tiders
	Nn-Indef-Sg	Indefinite singular noun	stol, lekam
	Nn-Indef-Sg-Gen	Genitive indefinite singular noun	fridoms
	Nn-Indef-SP	Indefinite singular or plural noun	lys, skritt
	Nn-Letter	Lowercase and uppercase letters	b, N
	Nn-Net	URL and e-mail address	www.inxight.com info@inxight.com
Num	Num	Cardinal numeric expression or plural cardinal number (spelled out)	-294, 4,6%, xii, 1.100 to, tre
	Num-Def-Sg	Definite singular cardinal number "one" (spelled out)	eine
	Num-Indef-Sg	Indefinite singular cardinal number (spelled out)	eitt
Ord	Ord	Ordinal number (in digits or spelled out)	7., første
Part	Part-Inf	Infinitival particle	**å** kalla

Umbrella Tag	Complete Tag	Description	Examples
Prep	Prep	Preposition	med, ut, opp
	Prep-av	Preposition **av**	av
	Prep-for	Preposition **for**	for
	Prep-fra	Preposition **frå**	**frå** sin opposisjon
	Prep-i	Preposition **i**	i
	Prep-paa	Preposition **på**	**på** alle
	Prep-ved	Preposition **ved**	ved
Pron	Pron-Acc	Accusative pronoun	henne
	Pron-Nom	Nominative pronoun	han, ho
	Pron-Poss-Pl	Possessive pronoun with plural agreement	sine
	Pron-Poss-Sg	Possessive pronoun with singular agreement	sin
Prop	Prop	Proper name	Johan
	Prop-Gen	Genitive proper name	Espens
Punct	Punct	Miscellaneous punctuation	- [>
	Punct-Comma	Comma	,
	Punct-Quote	Quotation marks	" ' ' << >> " "
	Punct-Sent	Sentence boundary punctuation	. ... ? : ; !

Umbrella Tag	Complete Tag	Description	Examples
V	V-Impv	Imperative verb	speil, kann
	V-Inf	Infinitive verb	gjera
	V-Inf-SForm	S-Form of infinitive verb	belastast, synast
	V-PaPart	Past participle verb	peika
	V-PaPart-SForm	S-Form of past partici-ple verb	trivest
	V-Past	Past tense verb	sa
	V-Past-SForm	S-Form of past tense verb	møttest, mistreivst, syntest
	V-Pres	Present tense verb	gir, oppfattar
	V-Pres-SForm	S-Form of present tense verb	finst
	V-PrPart	Present participle verb	seg **nemnande**

5.17.1.5 Grouping in Nynorsk

A Nynorsk simple noun phrase consists minimally of one or more nouns or proper nouns, for example:

- Johan Brox

Nouns are grouped with premodifying adjectives, possessives, and indefinite nouns, as in:

- anvendt forsking
- modernismens kris
- økonom Tormod Hermannsen

Nynorsk noun phrases can also include compound parts, and they can be coordinated with **og** and **eller**, for example:

- stats- og folkekyrkja
- kommunikasjon og inngangsport
- personane eller gruppene

A simple noun phrase also combines with prepositional phrases beginning with **av** and **frå**, as in:

- overtatt av staten
- betong frå sementfabrikken

Noun groups also include the prepositions **ved**, **i**, and **på** when they are followed by a proper noun.

- semesteropning ved Volda Lærarhøgskule
- redaktør i Fjeld-Ljom
- sosialkomiteen på Stortinget

Prepositional phrases starting with **for** are included when they follow proper nouns:

- Sundet for fulle segl

5.17.2 Extraction

This section describes the extraction-specific information for Nynorsk.

5.17.2.1 Predefined Entity Types

This section describes the predefined entity types supported by the Nynorsk language module and examples of each.

5.17.2.1.1 NOUN_GROUP

A Nynorsk simple noun phrase consists minimally of one or more nouns or proper nouns, for example:

- Johan Brox

Nouns are grouped with premodifying adjectives, possessives, and indefinite nouns, as in:

- anvendt forsking
- modernismens kris
- økonom Tormod Hermannsen

Nynorsk noun phrases can also include compound parts, and they can be coordinated with **og** and **eller**, for example:

- stats- og folkekyrkja
- kommunikasjon og inngangsport

- personane eller gruppene

5.18 Polish Language Reference

This chapter describes the behavior of the Polish language module.

5.18.1 Linguistic Processing

This section describes the language-specific information on the processing of Polish texts, including word segmentation and stemming.

5.18.1.1 Character Encodings for Polish

- iso_8859_2
- cp_1250
- utf_8, utf_16, ucs_4

5.18.1.2 Word Segmentation in Polish

The Polish segmenter follows all of the general segmentation rules in the white space languages.

Related Topics

- Word Segmentation

5.18.1.3 Stemming in Polish

The Polish stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Examples
Noun	hosannami -> hosanna, fyrgolu -> fyrgol
Verb	śle -> słać, zajętego -> zająć
Adjective	profonicznym -> profoniczny, progościnniejsze -> progościnny
Adverb	procale -> procały, wtyczkowie -> wtyczkowy

5.18.2 Extraction

Note:
Polish is a basic-level supported language, which means it supports extraction by using dictionaries or extraction rules only.

5.19 Portuguese Language Reference

This chapter describes the behavior of the Portuguese language module.

5.19.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Portuguese texts, including word segmentation and stemming.

5.19.1.1 Character Encodings for Portuguese

- iso_8859_1
- cp_1252
- utf_8, utf_16, ucs_4

5.19.1.2 Word Segmentation in Portuguese

The Portuguese segmenter follows all of the general segmentation rules in the white space languages. The Portuguese segmenter has the following language-specific behavior.

Clitics are not split off, and combined words are treated as one word.

Text	Segmented
dir-se-ia	dir-se-ia
pela	pela

Related Topics

- Word Segmentation

5.19.1.3 Stemming in Portuguese

This section describes the standard stemmer and the expanded inflectional stemmer used for stemming in Portuguese.

5.19.1.3.1 Standard Stemmer

The Portuguese stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Baseform	Examples
Noun	(Masculine) non-diminutive singular	filhos -> filho, balinha -> bala
Verb	Infinitive	traremos -> trazer, alimentará -> alimentar
Adjective	Masculine singular	bonitona -> bonito, caríssimos -> caro
Adverb	Positive form or source form	ultimamente -> ultimamente, pessimamente -> pessimamente, mal -> mal

Contracted prepositions and pronouns are broken into their component parts, and these stems are returned with an equal sign in between, indicating that the stems are of equal semantic importance. If the contracted preposition occurs in a multiword unit, then the final contraction is broken. This is shown in the following table:

Example	Stem
pelo	por=o
dele	de=ele
abaixo deste	abaixo de=este
ma	eu=ela

5.19.1.3.2 Expanded Inflectional Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text (such as email, online documents, or queries). See Expanded Inflectional Stemming for general behavior. . Here, we list the specifics for Portuguese.

The expanded version does not require correct capitalization and accentuation.

Case Variants

The expanded version accepts lower case letters in addition to capital letters for those words where the capitals are obligatory.

Example	Output
Varig	Varig
varig	Varig
USA	USA
usa	USA

Deaccented Characters

The expanded version also allows deaccented characters in place of accented ones.

Example	Output
mãos	mão
maos	mão
faríeis	fazer
farieis	fazer

Hyphenation

Hyphens in non-numeric expressions are optional in the expanded version.

Example	Output
Port-Royal	Port-Royal
PortRoyal	Port-Royal

5.19.1.4 Part-of-Speech Tagging in Portuguese

The following table shows the Portuguese tag set available for using in defining custom entities. The tag names are accompanied by a brief description and one or more examples. If the example consists of more than one word, the word exemplifying the current tag is in **bold**. The tag set makes no distinction for gender.

`Adj`	`Adj`	Invariant adjective	simples
	`Adj-Pl`	Plural adjective	cidãos **portugueses**
	`Adj-Sg`	Singular adjective	continente **europeu**
`Adv`	`Adv`	Adverb	directamente
	`Adv-Deg`	Adverbs that can modi-fy adjectives	**mais** livre

Aux	Aux-be	Auxiliaries of **ser** and **estar**	são, estão
	Aux-have	Auxiliaries of **ter** and **haver** ('have')	tem, haverá
	Aux-Inf-be	Infinitive auxiliary **estar**	estar
	Aux-Inf-have	Infinitive form of **ter** and **haver**	ter, haver
	Aux-Inf-Pron-be	Infinitive 'be' auxilliary with attached clitic	sê-lo
	Aux-Inf-Pron-have	Infinitive of **ter** and **haver** with attached clitic	ter-se
	Aux-Pron-be	'be' auxiliaries with attached clitic	estava-me
	Aux-Pron-have	Auxiliaries **ter** and **haver** with attached clitic	tinham-se
Conj	Conj	Unclassified conjunctions	nem, aquando, tal como
	Conj-Comp	Comparison conjunction	mais **do que** uma vez
	Conj-Coord	Coordinating conjunction	por fax **ou** correio
	Conj-Sub	Subordinating conjunction	para que, se, que

Det/Pron	Det/Pron-Dem-Pl	Plural demonstrative determiner or pronoun	estes, aqueles
	Det/Pron-Dem-Sg	Singular demonstrative determiner or pronoun	este, aquele
	Det/Pron-Poss-Pl	Plural possessive determiner or pronoun	vossos, seus
	Det/Pron-Poss-Sg	Singular possessive determiner or pronoun	vosso, seu
	Det/Pron-Quant-Pl	Plural quantifying determiner or pronoun	**quantas** vezes
	Det/Pron-Quant-Sg	Singular quantifying determiner or pronoun	**quanta** vez
Det	Det-Int	Interrogative determiner	demostra a **que** ponto
	Det-Int-Pl	Plural interrogative determiner	quantos, quantas, quais
	Det-Int-Sg	Singular interrogative determiner	quanto, quanta, qual
	Det-Pl	Plural determiner	**os** maiores aplausos
	Det-Rel-Pl	Plural relative determiner	cujas
	Det-Rel-Sg	Singular relative determiner	cuja
	Det-Sg	Singular determiner	**o** service
Interj	Interj	Interjection or onomatopoeia	oh, claro

Nn	Nn	Invariant noun	caos
	Nn-Letter	Lowercase and uppercase letters, by themselves or followed by a period or right parenthesis	b, N
	Nn-Net	URL and e-mail address	www.inxight.com info@inxight.com
	Nn-Pl	Plural noun	serviços
	Nn-Sg	Singular noun	esta **rede**
Num	Num	Numeric expression	123
Part	Part-Neg	Negation particle	nunca

Prep	Prep	Preposition	com
	Prep-a	Preposition **a**	a
	Prep-Adv	Combination preposition and adverb	venho **daqui**
	Prep-de	Preposition **de**	de
	Prep-Dem-Pl	Combination preposition and plural demonstrative	**desses** recursos
	Prep-Dem-Sg	Combination preposition and singular demonstrative	**nesta** placa
	Prep-Det-Pl	Combination preposition and plural determiner	nas, longe das
	Prep-Det-Pl-a	Combination **a** and plural determiner	aos
	Prep-Det-Pl-de	Combination **de** and pllural determiner	**dos** Grandes Bancos
	Prep-Det-Sg	Combination preposition and singular determiner	**na** construção
	Prep-Det-Sg-a	Combination **a** and singular determiner	ao
	Prep-Det-Sg-de	Combination **de** and singular determiner	da, doutro
	Prep-para	Preposition **para**	para
	Prep-Pron	Combination preposition and pronoun	atrás dela
	Prep-Quant-Pl	Combination preposition and plural quantifier	**nuns** terrenos
	Prep-Quant-Sg	Combination preposition and singular quantifier	**numa** nuvem
	Prep-Rel		nesta praia **aonde** ...

		Combination preposition and relative pronoun	
	Prep-Rel-Pl	Combination preposition and plural relative pronoun	alunos **aos quais**
	Prep-Rel-Sg	Combination preposition and singular relative pronoun	área através do qual
Pron	Pron	Invariant pronoun	si
	Pron-Int-Pl	Plural interrogative pronoun	**Quais** são os livros de Manuel?
	Pron-Int-Sg	Singular interrogative pronoun	**Qual** é o livro dela?
	Pron-Pl	Plural pronoun	eles
	Pron-Rel	Invariant relative pronoun	um ortopedista **que**
	Pron-Rel-Pl	Plural relative pronoun	as instalações **as quais**
	Pron-Rel-Sg	Singular relative pronoun	o ensayo **o qual**
	Pron-Sg	Singular pronoun	ele
Prop	Prop	Proper noun	Lisbon, Windows
Punct	Punct	Other punctuation	: ( )
	Punct-Comma	Comma	,
	Punct-Sent	Sentence punctuation	. ! ? ;
V/Adj	V/Adj-PaPart	Past participle verb or adjective	penetrado, referida

V	V-Fin	Finite verb	corresponde
	V-Fin-Pron	Finite verb with attached clitic	deu-lhe
	V-Inf	Infinitive verb	reunir, conservar
	V-Inf-Pron	Infinitive verb with attached clitic	datar-se
	V-PrPart	Present participle verb	falando
	V-PrPart-Pron	Present participle verb with attached clitic	deixando-a

5.19.1.5 Grouping in Portuguese

A Portuguese simple noun phrase is a noun with optional pre- and postmodifiers. Nouns may be preceeded by an adjective, as in:

- diferentes destinos

Postmodifiers include adjectives and nouns, for example:

- água salgada
- Monte Sinai

Nouns are also grouped with a following prepositional phrase beginning with **de** and including a (possibly modified) noun:

- mastro de emergência

5.19.2 Extraction

This section describes the extraction-specific information for Portuguese.

5.19.2.1 Predefined Entity Types

This section describes the predefined entity types supported by the Portuguese language module and examples of each.

5.19.2.1.1 NOUN_GROUP

A Portuguese simple noun phrase is a noun with optional pre- and postmodifiers. Nouns may be preceeded by an adjective, as in:

- diferentes destinos

Postmodifiers include adjectives and nouns, for example:

- água salgada
- Monte Sinai

5.20 Romanian Language Reference

This chapter describes the behavior of the Romanian language module.

5.20.1 Linguistic Processing

This section describes the language-specific information on the processing of Romanian texts, including word segmentation and stemming.

5.20.1.1 Character Encodings for Romanian

- iso_8859_2
- cp_1250
- utf_8, utf_16, ucs_4

5.20.1.2 Word Segmentation in Romanian

The Romanian segmenter follows all of the general segmentation rules in the white space languages.

Related Topics

- Word Segmentation

5.20.1.3 Stemming in Romanian

Stemming in Romanian includes the standard stemmer and the expanded stemmer.

5.20.1.3.1 Standard Stemmer

The Romanian stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Examples
Noun	profesorul -> profesor, muzeele -> muzeu, marii -> mare
Verb	terminam -> termina, doresc -> dori, credeam -> crede
Adjective	frumoasa -> frumos, mici -> mic, eficace -> efi-cace
Adverb	aici -> aici, teoretic -> teoretic, mai -> mai

5.20.1.3.2 Expanded Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text, such as email, online documents, or queries. In Romanian, this includes accented characters missing their diacritics and proper names without word-initial capitalization. For instance:

Case Variants

The expanded version accepts lower case letters in addition to capital letters for those words where the capitals are obligatory.

Example	Output
România	România
România	România

Deaccented Characters

The expanded version also allows deaccented characters in place of accented ones.

Example	Output
inimă	inimă
inima	inimă

5.20.2 Extraction

Note:
Romanian is a basic-level supported language module, which means it supports extraction by using dictionaries or extraction rules only.

5.21 Russian Language Reference

This chapter describes the behavior of the Russian language module.

5.21.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Russian texts, including word segmentation and stemming.

5.21.1.1 Character Encodings for Russian

- iso_8859_5
- cp_1251
- koi8_r
- utf_8, utf_16, ucs_4

5.21.1.2 Word Segmentation in Russian

The Russian segmenter follows all of the general segmentation rules in the white space languages. The Russian segmenter handles multiword units such as вряд ли and 4ръЧ фев., 07ОШб, as well as abbreviations like лаб. and фр .

Related Topics

- Word Segmentation

5.21.1.3 Stemming in Russian

The Russian stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Examples
Noun	работу -> работа, изменения -> изменение
Verb	покупаю -> покупать, едешь -> ехать покупала -> покупать, ехали -> ехать
Adjective	красного -> красный, краснее -> красный, краснейшим -> красный
Adverb	хорошо -> хорошо, ясно -> ясно

5.21.1.4 Part-of-Speech Tagging in Russian

The following table shows the Russian tag set. The tag names are accompanied by a brief description and one or more examples. The tag set makes no distinction for number or gender.

Umbrella Tag	Complete Tag	Description	Examples
Adj	`Adj-Nom`	Adjective in nominative case	красивый, красивая, красивое, красивые
	`Adj-Acc`	Adjective in accusative case	красивого, красивую, красивое, красивые
	`Adj-Gen`	Adjective in genitive case	красивого, красивой, красивых
	`Adj-Obl`	Adjective in oblique case (dative, instrumental, and so on.)	красивым, красивой, красивому, красивыми
	`Adj-Comp`	Comparative adjective	краше
	`Adj-Brf`	Adjective in brief form	красив, красива, красивы
	`Adj-Inv`	Abbreviated adjective	т.н.

Umbrella Tag	Complete Tag	Description	Examples
Adv	Adv	Adverb	быстро
	Adv-Comp	Comparative adverb	лучше
Conj	Conj	Conjunction	и, но, чтобы
Det	Det-Nom	Pronominal adjective in nominative case	этот
	Det-Acc	Pronominal adjective in accusative case	эту
	Det-Gen	Pronominal adjective in genitive case	нашей
	Det-Obl	Pronominal adjective in oblique case	этому
	Det-Inv	Abbreviated pronominal adjective	др.
Dig	Dig	Number (in digits)	1999, 100Мб
Interj	Interj	Interjection	ага, ах, ба
Nn	Nn-Nom	Noun in nominative case	сестра, сестры
	Nn-Acc	Noun in accusative case	сестру, сестер
	Nn-Gen	Noun in genitive case	сестер
	Nn-Obl	Noun in oblique case	сестрой, сестрами
	Nn-Inv	Abbreviated noun	пр., о., г.
Num	Num	Number	три, восемь
Ord	Ord	Ordinal number (in digits)	7., 3.

Umbrella Tag	Complete Tag	Description	Examples
Pron	Pron-IntRel-Nom	Relative pronoun in nominative case	кто
	Pron-IntRel-Acc	Relative pronoun in accusative case	кого
	Pron-IntRel-Gen	Relative pronoun in genitive case	чего
	Pron-IntRel-Obl	Relative pronoun in oblique case	кому
	Pron-Pers-Nom	Personal prounoun in nominative case	я, ты
	Pron-Pers-Acc	Personal prounoun in accusative case	меня, тебя
	Pron-Pers-Gen	Personal prounoun in genitive case	меня, тебя
	Pron-Pers-Obl	Personal prounoun in oblique case	мной, тобой
	Pron-Adv	Pronominal adverb	откуда, кое-как
	Pron-Nom	Prounoun in nominative case	все, ничто
	Pron-Acc	Prounoun in accusative case	все
	Pron-Gen	Prounoun in genitive case	всего, ничего
	Pron-Obl	Prounoun in oblique case	всеми, ничем
Prep	Prep-Nom	Preposition governing nominative case	плюс, минус
	Prep-Acc	Preposition governing accusative case	за
	Prep-Gen	Preposition governing genitive case	без, накануне
	Prep-Obl	Preposition governing oblique case	благодаря, к

Umbrella Tag	Complete Tag	Description	Examples
Prop	Prop-Nom	Proper name in nominative case	Москва, Мальцев
	Prop-Acc	Proper name in accusative case	Москву
	Prop-Gen	Proper name in genitive case	Москвы
	Prop-Obl	Proper name in oblique case	Москве, Мальцеве
Punct	Punct-Comma	Comma	,
	Punct-Sent	Punctuation symbol at the end of a sentence	. ? !
	Punct-Symbol	Any separator in a sentence	% / $
Part	Part	Particle	аж, же
	Part-Int	Introduction particle	авось
	Part-Sent	Sentence particle	аминь
	Part-Mood	Mood marker particle	бы, ли
Aux	Aux	Auxiliary verb	быть

Umbrella Tag	Complete Tag	Description	Examples
Verb	Verb-Fin	Finite verb	делай, делает, делал
	Verb-Ger	Adverbial participle (gerund)	делав, делавши, делая
	Verb-Inf	Infinitive verb	делать
	Verb-Acc	Participle in accusative case	делавшего, делавшую
	Verb-Gen	Participle in genitive case	делавшего, делавшей
	Verb-Nom	Participle in nominative case	делавший, делавшее, делавшая
	Verb-Obl	Participle in oblique case	делавшим, делавшей
	Verb-Brf	Participle in brief form	делано, делана
	Verb-Inv	Abbreviated Verb	исп.

5.21.2 Extraction

This section describes the extraction-specific information for Russian.

5.21.2.1 Russian Subtypes

Russian supports subtypes in the types `ORGANIZATION` and `URI`.

Related Topics

- Subtypes

5.21.2.2 Predefined Entity Types

This section describes the predefined entity types supported by the Russian language module and examples of each. Click each link to jump to that subsection: COUNTRY, GEO_AREA, GEO_FEATURE, LOCALITY, ORGANIZATION, PERSON, PHONE, PROP_MISC, TITLE, and URI

5.21.2.2.1 COUNTRY

Names of countries, including abbreviations:

- Германия
- Россия
- Северная Корея
- США
- ОАЭ

5.21.2.2.2 GEO_AREA

A geographical area larger than a city that captures a significant land mass and embeds a set of cities and towns such as continent, group of countries, state, autonomous region, and so on:

- Центральная Европа
- Якутия
- Огайо
- штат Уттар-Прадеш
- республика Бурятия
- Сахалинская область
- Дальний Восток
- Ненецкий автономный округ

5.21.2.2.3 GEO_FEATURE

Names of districts, small towns, and villages, or rivers, lakes and mountains:

- Чистопольский район
- поселок Кутопьюган
- озеро Иссык-Куль
- Каспийское море

- река Волга

5.21.2.2.4 LOCALITY

City names:

- Таганрог
- Нью-Йорк
- Берлин
- Великие Луки
- Санкт-Петербург

City names preceded by a directional designator:

- северо-запад Москвы
- восток Лондона

5.21.2.2.5 ORGANIZATION

Government, legal, and service agencies, including non-profit organizations, fine arts groups, and other associations and institutions, extracted as one of the following subtypes:

- `COMMERCIAL`–The name of commercial organizations, such as major companies or corporations. For example:
  - АФК Система
  - Майкрософт
  - Газпром
  - Газпром-Медиа
  - МТС
  - Норникель
  - компания "СМВБ-Информационные технологии"
  - ЗАО “Эдем”
  - холдинг "Телекоминвест"
  - Автобанк-Никойл
  - Автомобильный Банкирский Дом
  - Росевробанк
  - Банк Москвы

Note:
Any extracted `ORGANIZATION/COMMERCIAL` entities can be parsed and standardized using the Data Quality Data Cleanse transform by mapping them to one of the `FIRM` input fields.

- `OTHER`–Any organization that does not fit into a more specific subtype:
  - правительство США
  - Международная организация по стандартизации
  - Санкт-Петербургский государственный университет
  - ООН
  - Ассоциация независимых центров экономического анализа
  - Министерство по налогам и сборам

5.21.2.2.6 PERSON

Variations of names:

- Валерий Трошин
- Наталья Фотиева
- Алексей Иванович Сергеев
- П.В. Шавенков
- Иван
- Иван Иванович
- Сокуров
- Джордж Буш-младший
- Фритц Фидлер
- Татьяна Щепкина-Куперник
- Суворова Мария Георгиевна

5.21.2.2.7 PHONE

Russian and international phone numbers:

- 8(920) 284 8484
- (+7495) 771 7226
- +7(495)788-97-99

Coordinated chains of phone numbers:

- телефоны 2100500 или 2222222

- тел.: (8-0512)-21-81-60, 49-21-92, 47-88-97

5.21.2.2.8 PROP_MISC

A proper name that does not fall into any of the entity types specified by the other entities:

- В финале последнего Кубка Кремля теннисистка добилась победы
- Заодно можно проверить готовность города к Олимпиаде 2008
- Между тем, согласно исследованию "Аэртон", депозиты в долларах показали отрицательную доходность
- В годы Второй Мировой войны офицер был капитаном жандармерии

5.21.2.2.9 TITLE

The description of a person's position. The position entity contains a complement expressed by organization, company or geographical name:

- генеральный директор Агентства прикладной и региональной политики
- генеральный секретарь ОПЕК
- председатель Ассоциации коммуникационных агентств России
- ректор Военно-медицинской академии
- глава Генеральной прокуратуры
- президент Франции
- адвокат экс-главы "ЮКОСа"

5.21.2.2.10 URI

An address on the internet:

- www.yandex.ru
- http://blog.kp.ru/community/1231628
- erkki-vahamaa@kajaani.fi

5.22 Serbian Language Reference

This chapter describes the behavior of the Serbian language module.

5.22.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Serbian texts, including word segmentation and stemming.

5.22.1.1 Character Encodings for Serbian

- iso_8859_2
- cp_1250
- utf_8, utf_16, ucs_4

5.22.1.2 Word Segmentation in Serbian

The Serbian segmenter follows all of the general segmentation rules in the white space languages.

Related Topics

- Word Segmentation

5.22.1.3 Stemming in Serbian

This section describes the standard stemmer and the expanded stemmer used for stemming in Serbian.

5.22.1.3.1 Standard Stemmer

The standard Serbian stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Examples
Noun	nemanja -> nemanje , teglu -> tegla , odgovorom -> odgovor
Verb	donese -> doneti , ponude -> ponuditi , zadovolji-mo -> zadovoljiti
Adjective	srbijansku -> srbijanski , spremni -> spreman , izborni -> izboran
Adverb	joj -> ona , to -> taj , neku -> neki

5.22.1.3.2 Expanded Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text, such as email, online documents, or queries. In Serbian, this includes accented characters missing their diacritics and proper names without word-initial capitalization.

Case Variants

The expanded version accepts lower case letters in addition to capital letters for those words where the capitals are obligatory. Initial capital letters will be lowercased to deal with those cases where common nouns are capitalized, for example, Informacione Tehnologije .

Example	Output
Srbija	Srbija
srbija	Srbija
Plovka	plovka
plovka	plovka
Splet	splet
splet	splet

Characters with Missing Diacritics

The expanded version also allows characters with missing diacritics in place of characters with diacritics. Foe example,

Example	Output
bajačica	bajačica
bajacica	bajačica

5.22.1.4 Part-of-Speech Tagging in Serbian

The following table shows the Serbian tag set available for using in defining custom entities. The tag names are accompanied by a brief description and one or more examples. The tag set makes no distinction for gender.

Umbrella Tag	Complete Tag	Description	Examples
`Abbr`	`Abbr`	Abbreviation	napr
`Adj`	`Adj`	Adjectives	mnogima, srpskog
	`Adj-Comp`	Comparative adjectives	dublji, manje
	`Adj-Sup`	Superlative adjectives	najnovija, najgore
`Adv`	`Adv`	Adverbs	bar, tako
	`Adv-Comp`	Comparative adverb	bolje, smelije
`Conj`	`Conj`	Conjunctions	da, zato
`Enum`	`Enum`	Enumeration	etc.
`Interj`	`Interj`	Interjection	ne, li

Umbrella Tag	Complete Tag	Description	Examples
Nn	Nn-Pl-Nom	Plural nominative noun	srbi, ljudi
	Nn-Pl-Acc	Plural accusative noun	gorštačkim, razloge
	Nn-Pl-Gen	Plural genitive noun	svetinja, vekova
	Nn-Pl-Case	Plural, including vocative, locative and instrumental noun	funkcionerima, uslovima
	Nn-Sg-Nom	Singular nominative noun	istina, pravda
	Nn-Sg-Acc	Singular accusative noun	put, narod
	Nn-Sg-Gen	Singular, genitive noun	godine, poverenja
	Nn-Sg-Case	Singular, including vocative, locative and instrumental noun	ratu, knjizi
Num	Num	Digits	123
	Num-Nom	Nominative number expression	desetoro
	Num-Acc	Accusative number expression	dvoje
	Num-Case	Number expression other than nominative and accusative	troje
	Num-Card	Cardinal number	jedan, devet
	Num-Ord	Ordinal number	prvu, osmo
Prep	Prep	Preposition	za, od

Umbrella Tag	Complete Tag	Description	Examples
Pron	Pron	Pronoun	svog, te
	Pron-Pl	Plural pronoun	koje
	Pron-Sg	Singular pronoun	šta
	Pron-Ref	Reflexive pronoun	se
	Pron-Pers-Sg	Singular personal pronoun	mi
	Pron-Pers-Pl	Plural personal pronoun	ih
	Pron-Poss-Sg	Singular possesive pronoun	našoj
	Pron-Poss-Pl	Plural possessive pronoun	njegovih
Prop	Prop	Proper name	Zagreb
Punct	Punct-Sent	Sentence ending punctuation	! ? .
	Punct-Comma	Comma	,
	Punct-Open	Opening punctuation	(
	Punct-Close	Closing punctuation	)
	Punct	Other punctuation	…
V	V-Inf	Infinitive verb	objasniti, uništiti
	V-Fin-Sg	Singular finite verb	reci
	V-Fin-Pl	Plural finite verb	smatraju, istaknemo
	V-Part	Participle	izvadivši
	V-Part-Sg	Singular participle	napao, dozvolio
	V-Part-Pl	Plural participle	iskopali, proganjali
	V-Aux-Clit	Auxiliary verb	nisu, bi

5.22.1.4.1 Unfound Words

Words not found in the tagger dictionary are passed to the Serbian guesser to be assigned the most likely tag. The Serbian guesser assigns tags to unfound words based on a set of rules about Serbian morphology. For example, a word ending in `sti –»` is likely an infinitive verb. Internet and e-mail addresses are assigned the tag `Nn-Net`. Capitalization information is also important; for instance, capitalized words tend to be guessed as proper nouns.

5.22.1.5 Grouping in Serbian

Serbian noun groups can be simple or compound nouns with their modifiers. Modifiers can be adjectives, adjectival pronouns, or ordinal numbers but not determiners, personal, or **wh-** pronouns. Modifiers can have adverbs as their own modifiers. For example:

- dole podpisani pravoslavni srpski sveštenici
- hiljadugodišnjim iskustvom
- posle srpskih seoba

5.22.2 Extraction

This section describes the extraction-specific information for Serbian.

5.22.2.1 Predefined Entity Types

This section describes the predefined entity types supported by the Serbian language module and examples of each.

5.22.2.1.1 NOUN_GROUP

Serbian noun groups can consist of one or more nouns with optional premodifiers.

The premodifiers can consist of zero or more adverbs followed by one or more (possibly coordinated) adjectives, adjectival pronouns, or adjectival numerals.

For example:

- poslednje ostatke Krsta
- duhu politike
- petvekovno tursko ropstvo

5.23 Slovak Language Reference

This chapter describes the behavior of the Slovak language module.

5.23.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Slovak texts, including word segmentation and stemming.

5.23.1.1 Character Encodings for Slovak

- iso_8859_2
- cp_1250
- utf_8, utf_16, ucs_4

5.23.1.2 Word Segmentation in Slovak

The Slovak segmenter follows all of the general segmentation rules in the white space languages.

Related Topics

- Word Segmentation

5.23.1.3 Stemming in Slovak

This section describes the standard stemmer and the expanded stemmer used for stemming in Slovak.

5.23.1.3.1 Standard Stemmer

The standard Slovak stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Examples
Noun	jablká -> jablko, brány -> brána, domom -> dom, stoly -> stôl
Verb	chcel -> chcit', prosím -> prosit', boli -> byt', myslí -> myslieť
Adjective	tmavom -> tmavý, úzkej -> úzký, starąí -> starý
Adverb	dobre -> dobre, nikde -> nikde, neskôr -> neskôr

5.23.1.3.2 Expanded Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text, such as email, online documents, or queries. In Slovak, this includes accented characters missing their diacritics and proper names without word-initial capitalization.

Case Variants

The expanded version accepts lower case letters in addition to capital letters for those words where the capitals are obligatory. Initial capital letters will be lowercased to deal with those cases where common nouns are capitalized, for example, Stredoveká Literatúra.

Example	Output
Bratislava	Bratislava

Example	Output
bratislava	Bratislava
Ide	Ida, idea, ísť
ide	Ida, idea, ísť
Literatúra	literatúra
literatúra	literatúra

Deaccented Characters

The expanded version also allows deaccented characters in place of accented ones.

Example	Output
lačný	lačný
lacny	lačný

5.23.1.4 Part-of-Speech Tagging in Slovak

The following table shows the Slovak tag set available for using in defining custom entities. The tag names are accompanied by a brief description and one or more examples. The tag set makes no distinction for gender.

The following table shows the Slovak tag set available for using in defining custom entities. The tag names are accompanied by a brief description and one or more examples. The tag set makes no distinction for gender.

Umbrella Tag	Complete Tag	Description	Examples
`Abbr`	`Abbr`	Abbreviation	dopr., hl

Umbrella Tag	Complete Tag	Description	Examples
Adj	Adj	Adjectives	úškrnových, úšustov
	Adj-Comp	Comparative adjectives	účtovovanší, účtovovanším
	Adj-Sup	Superlative adjectives	najúbohším, najúbohších
Adv	Adv	Adverbs	najavo, sami
	Adv-Comp	Comparative adverbs	účastnšie, účastnejšie
	Adv-Conj	Either adverb or conjunction	ako, kde, tak
	Adv-Part	Either adverb or particle	celkom, ešte, práve
	Adv-Sup	Superlative adverbs	najúbohšie, najúlisnšie
Conj	Conj	Conjunction	alebo, keby, pritom
	Conj-Part	Either conjunction or particle	a, aj, ale
Interj	Interj	Interjection	zbohom, výborne
Nn	Nn	Invariant noun	zombi, šapitó
	Nn-Pl-Gen	Plural, genitive noun	účtov, účtovaní
	Nn-Pl-Case	Plural, nominative, accusative, dative, locative and instrumental noun	účtami, účtovaniami
	Nn-Sg-Gen	Singular, genitive noun	účtu, účtovania
	Nn-Sg-Case	Singular, nominative, accusative, dative, locative and instrumental noun	účtovi, účtovaním
	Nn-Net	URL, e-mail address	www.inxight.com info@inxight.com

Umbrella Tag	Complete Tag	Description	Examples
Num	Num	Number expression other than cardinal or ordinal, ascii numbers	1, 12%
	Num-Card	Cardinal number	osemsto, štyritisíc osemsto
	Num-Ord	Ordinal number	dvetisíc, dvetisícsto
Part	Part	Particles	nie, by
Pref	Pref	Prefix (stand alone prefix)	vodo, ne
Prep	Prep	Prepositions	v, zo
Pron	Pron-Dem-Pl	Plural demonstrative pronoun	všelitakí, všelitakým
	Pron-Dem-Sg	Singular demonstrative pronoun	taký, všelitakom
	Pron	Indefinite pronoun	čosi
	Pron-Pl	Plural pronoun	dačíchsi, čiesi
	Pron-Sg	Singular pronoun	kdečiasi, všeličiasi
	Pron-Interrog	Interrogative pronoun	kto, všelikoho
	Pron-Refl	Reflexive pronoun	sám, svoj
	Pron-Pers-Sg	Singular personal pronoun	ona, on
	Pron-Pers-Pl	Plural personal pronoun	oni, ony
	Pron-Poss	Possesive pronoun	váš, ich
Prop	Prop	Prop	Swisscom, Swisscomami

Umbrella Tag	Complete Tag	Description	Examples
Punct	Punct-Sent	Sentence ending punctuation	! ? .
	Punct-Comma	Comma	,
	Punct-Open	Opening punctuation	(
	Punct-Close	Closing punctuation	)
	Punct-Quote	Quote	"
	Punct	Other punctuation	... -
V	V-Inf	Infinitive verb	účtovat, útočiet
	V-Past-Pl	Plural, past tense verb	účtovali, účinkovali
	V-Past-Sg	Singular, past tense verb	účtoval, účtovala
	V-Pres-Pl	Plural, present tense verb	účtovujú, účtujeme
	V-Pres-Sg	Singular, present tense verb	účtovujem, účtovuj
	V-Fut-Pl	Plural, future tense verb	budú, budete
	V-Fut-Sg	Singular, future tense verb	bude, budeš
	V-Aux	Auxiliary verb	vie, vieš

5.23.1.4.1 Unfound Words

Words not found in the tagger dictionary are passed to the Slovak guesser to be assigned the most likely tag. The Slovak guesser assigns tags to unfound words based on a set of rules about Slovak morphology. For example, a word ending in –» is likely an infinitive verb. Internet and e-mail addresses are assigned the tag `Nn-Net`.

Capitalization information is also important; for instance, capitalized words tend to be guessed as proper nouns.

5.23.1.5 Grouping in Slovak

Slovak noun groups can be simple or compound nouns with their modifiers.

Modifiers can be adjectives or ordinal numbers but not determiners or pronouns.

Modifiers can have adverbs as their own modifiers.

For example:

- rokovaniach orgánov Európskej únie
- základe poverenia poslancov Národnej rady

5.23.2 Extraction

This section describes the extraction-specific information for Slovak.

5.23.2.1 Predefined Entity Types

This section describes the predefined entity types supported by the Slovak language module and examples of each.

5.23.2.1.1 NOUN_GROUP

Slovak noun groups can be simple or compound nouns with their modifiers. Modifiers can be adjectives or ordinal numbers but not determiners or pronouns. Modifiers can have adverbs as their own modifiers.

For example:

- rokovaniach orgánov Európskej únie
- základe poverenia poslancov Národnej rady

5.24 Slovenian Language Reference

This chapter describes the behavior of the Slovenian language module.

5.24.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Slovenian texts, including word segmentation and stemming.

5.24.1.1 Character Encodings for Slovenian

- iso_8859_2
- cp_1250
- utf_8, utf_16, ucs_4

5.24.1.2 Word Segmentation in Slovenian

The Slovenian segmenter follows all of the general segmentation rules in the white space languages.

Related Topics

- Word Segmentation

5.24.1.3 Stemming in Slovenian

This section describes the standard stemmer and the expanded stemmer used for stemming in Slovenian.

5.24.1.3.1 Standard Stemmer

The standard Slovenian stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Examples
Noun	čtiva -> čtivo, čtivu -> čtivo, čbeličarju -> čbeličar, čvrstost -> čvrstost, čvrstostih -> čvrstost
Verb	jva -> jesti, jta -> jesti, jte -> jesti, je -> jesti, jesla -> jesti
Adjective	yorški -> yorški, yorških -> yorški, yorška -> yorški
Adverb	čvrsto -> čvrsto

5.24.1.3.2 Expanded Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text, such as email, online documents, or queries. In Slovenian, this includes accented characters missing their diacritics and proper names without word-initial capitalization.

Case Variants

The expanded version accepts lower case letters in addition to capital letters for those words where the capitals are obligatory. Initial capital letters will be lowercased to deal with those cases where common nouns are capitalized, for example, Informacijska Tehnologija.

Example	Output
Čile	Čile, čil
čile	Čile, čil
Tomaž	Tomaž
tomaž	Tomaž
Tehnologija	tehnologija
tehnologija	tehnologija

Deaccented Characters

The expanded version also allows deaccented characters in place of accented ones.

Example	Output
Čile	Čile, čil
Cile	Čile, čil

5.24.1.4 Part-of-Speech Tagging in Slovenian

The following table shows the Slovenian tag set available for using in defining custom entities. The tag names are accompanied by a brief description and one or more examples. The tag set makes no distinction for gender.

Umbrella Tag	Complete Tag	Description	Examples
`Adj`	`Adj`	Adjectives	miren, mirna, mlad, mladi
	`Adj-Comp`	Comparative adjectives	lepši, lepše, bolj divji, manj divji
	`Adj-Sup`	Superlative adjectives	najlepši, najlepše, najbolj divji, najmanj divji
`Adv`	`Adv`	Adverbs	lepo, naglo, nagloma
	`Adv-Comp`	Comparative adverbs	lepše, bolj vroče, manj razločno
	`Adv-Sup`	Superlative adverbs	najlepše, najbolj vroče, najmanj razločno
`Conj`	`Conj`	Conjunction	in, pa, medtem ko
	`Conj-Part`	Conjunction or participle	samo, ne
`Interj`	`Interj`	Interjection	pfuj, ehej

Umbrella Tag	Complete Tag	Description	Examples
Nn	Nn	Invariant noun, including abbreviation, acronyms and etc.	ZDA, št.
	Nn-Du-Gen	Dual, genitive noun	čvrstosti
	Nn-Du-Case	Dual, nominative, accusative, dative, locative and instrumental noun	čvrstostih, čvrstostima
	Nn-Pl-Gen	Plural, genitive noun	čvrstosti
	Nn-Pl-Case	Plural, nominative, accusative, dative, locative and instrumental noun	čvrstostih, čvrstostmi
	Nn-Sg-Gen	Singular,genitive noun	čvrstosti
	Nn-Sg-Case	Singular, nominative, accusative, dative, locative and instrumental noun	cvrstosti, čvrstostjo
	Nn-Net	URL, e-mail address	www.inxight.com info@inxight.com

Umbrella Tag	Complete Tag	Description	Examples
Num	Num	Invariant number expression	1, 12%
	Num-Card-Gen	Cardinal number, genitive	stotih
	Num-Card-Case	Cardinal number, nominative, assucative, dative, locative and instrumental	sto, stotim
	Num-Ord-Gen	Ordinal number, genitive	stotih
	Num-Ord-Case	Ordinal number, nominative, accusative, dative, locative and instrumental	sto, stotim
	Num-Gen	Number expressions other than cardinal or ordinal numbers, genitive	četvorke, četvork
	Num-Case	Number expressions other than cardinal or ordinal, nominative, accusative, dative, locative and instrumental	četvorka, četvorki
Part	Part	Particles	že, žal
Prep	Prep	Preposition	pod, po
	Prep-Cmpd	Preposition with clitic	podnje, podnjo

Umbrella Tag	Complete Tag	Description	Examples
Pron	Pron-Dem-Du	Dual demonstrative pronoun	toliki, tolikima
	Pron-Dem-Pl	Plural demonstrative pronoun	tolike, tolikimi
	Pron-Dem-Sg	Singular demonstrative pronoun	to, toliko
	Pron-Ref	Reflexive pronoun, invariant in number	sebe, seboj, sebi
	Pron-Ref-Sg	Singular reflexive pronoun	svoj, svojim, svoji
	Pron-Ref-Du	Dual reflexive pronoun	svoji, svojih, svojima
	Pron-Ref-Pl	Plural reflexive pronoun	svoji, svoje, svoja
	Pron-Pers-Sg	Singular personal pronoun	jaz, ti, on, ona, ono
	Pron-Pers-Du	Dual personal pronoun	midva, vidva, onadva
	Pron-Pers-Pl	Personal pronoun, plural	mi, me, vi, ve, oni
	Pron-Poss-Sg	Possesive pronoun, singular	moj, tvoj, njen, njegov
	Pron-Poss-Du	Possessive pronoun, dual	najin, vajin, njun
	Pron-Poss-Pl	Possessive pronoun, plural	naš, vaš, njihov
	Pron-Interrog	Interrogative pronoun	kdo, kaj, kateri
	Pron-Rel	Relative pronoun	kdor, kar, kateri, ki
	Pron-Pl	Plural pronoun	vsem, vsemi, vse, vsa
	Pron-Du	Dual pronoun	vsi, vsema
	Pron-Sg	Singular pronoun	vso, vsm, vse
	Pron	Other pronouns, indefinite, evaluative and etc.	isti, drug
Prop	Prop	Proper noun	Sava, Ljubljana Prop

Umbrella Tag	Complete Tag	Description	Examples
Punct	Punct-Sent	Sentence-ending punctuation	. ! ?
	Punct-Comma	Comma	,
	Punct-Open	Opening punctuation	(
	Punct-Close	Closing punctuation	)
	Punct-Quote	Quote	"
	Punct	Other punctuation	... -
V	V-Aux	Auxiliary verb	biti, bi
	V-Sup	Supine verb	prodat, spat
	V-Inf	Infinitive verb	prodati, spati
	V-PPast-Du	Dual, past tense verb	čvrcali, čvrčala
	V-PPast-Pl	Plural, past tense verb	čvrčala, čvrčale
	V-PPast-Sg	Singular, past tense verb	čvrčalo, čvrčal
	V-Pres-Du	Dual, present tense verb	jva, jta
	V-Pres-Pl	Plural, present tense verb	jte, jmo
	V-Pres-Sg	Singular, present tense verb	je, ješ

5.24.1.4.1 Unfound Words

Words not found in the tagger dictionary are passed to the Slovenian guesser to be assigned the most likely tag. The Slovenian guesser assigns tags to unfound words based on a set of rules about Slovenian morphology. For example, a word ending in **-ti** is likely an infinitive verb. Internet and e-mail addresses are assigned the tag `Nn-Net`.

Capitalization information is also important; for instance, capitalized words tend to be guessed as proper nouns.

5.24.1.5 Grouping in Slovenian

Slovenian noun groups are nouns with their modifiers.

Modifiers can be adjectives or ordinal numbers but not determiners or pronouns.

Modifiers can have adverbs as their own modifiers.

For example:

- življenju otrok
- Evropski uniji

5.24.2 Extraction

This section describes the extraction-specific information for Slovenian.

5.24.2.1 Predefined Entity Types

This section describes the predefined entity types supported by the Slovenian language module and examples of each.

5.24.2.1.1 NOUN_GROUP

Slovenian noun groups can be simple or compound nouns with their modifiers. Modifiers can be adjectives or ordinal numbers but not determiners or pronouns. Modifiers can have adverbs as their own modifiers.

For example:

- življenju otrok
- Evropski uniji

5.25 Spanish Language Reference

This chapter describes the behavior of the Spanish language module.

5.25.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Spanish texts, including word segmentation and stemming.

5.25.1.1 Character Encodings for Spanish

- iso_8859_1
- cp_1252
- utf_8, utf_16, ucs_4

5.25.1.2 Word Segmentation in Spanish

The Spanish segmenter follows all of the general segmentation rules in the white space languages. The Spanish segmenter has the following language-specific behavior.

Contracted words such as del or al are not split. Clitics are not separated, for example, in dámelo. Trailing hyphens are split apart from their words. Ordinal numbers are not separated from their period.

Text	Segmented
dámelo	dámelo
del	del

Text	Segmented
empresa-	empresa
	-
2a.	2a.

Related Topics

- Word Segmentation
- White Space Languages

5.25.1.3 Stemming in Spanish

This section describes the standard stemmer, the expanded inflectional stemmer, and the inflectional stemmer guesser used for stemming in Spanish.

5.25.1.3.1 Standard Stemmer

The Spanish stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. Proper nouns stem to themselves. All diminutive endings are removed, even in given names. This is shown in the table below.

Category	Baseform	Examples
Noun	Non-diminutive singular	caballitos -> caballo; gatos -> gato
Proper Noun	Non-diminutive source form	África -> África; Anita -> Ana
Verb	Infinitive	compuesto -> componer; contés-tame-> contestar
Adjective	Masculine singular	altas -> alto; chiquito -> chico

Category	Baseform	Examples
Adverb	Source form	por qué -> por qué; cariñosa-mente -> cariñosamente

Spanish pronouns are stemmed in the following way. All uninflecting forms stem to themselves. Plural-only forms and all personal pronouns maintain their number information. If applicable, these pronouns are stemmed to the nominative form. All other forms stem to the masculine, singular form. This is shown in the table below:

Text	Stem
algo	algo
ambas	ambos
ellas	ellos
mí	yo
éstas	éste

Closed class words like determiners and ordinal numbers are stemmed to the masculine, singular, nominative form. Non-inflecting word categories stem to themselves, for example, conjunctions, cardinal numbers and prepositions.

Text	Stem
esta	este
con	con

Acronyms, abbreviations and multiword units stem to themselves. Pronoun abbreviations stem to their full form. These behaviors are shown in the following table:

Text	Stem
UNAM	UNAM
p.ej.	p.ej.
Ud.	usted
los tuyos	el tuyo

Contracted words are stemmed into their component parts:

Text	Stem
conmigo	con=yo
al	a=el

5.25.1.3.2 Expanded Inflectional Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text (such as email, online documents, or queries). See Expanded Inflectional Stemming for the general behavior. The specifics for Spanish follow.

Hyphenation

The expanded version accepts optional hyphenation for words which normally have an obligatory hyphen.

Example	Output
MS-DOS	MS-DOS
MSDOS	MS-DOS
Baden-Baden	Baden-Baden

Example	Output
BadenBaden	Baden-Baden

Case Variants

The expanded version accepts lower case letters in addition to capital letters for those words where the capitals are obligatory.

Example	Output
OEA	OEA
oea	OEA

Deaccented Characters

The expanded version accepts completely deaccented characters in addition to accented ones.

Example	Output
corazón	corazón
corazon	corazón

5.25.1.3.3 Inflectional Stemmer Guesser

The inflectional stemmer guesser contains a set of morpological rules that you can apply to words that are unknown to the standard or expanded inflectional stemmer and therefore cannot be stemmed.

Linguistics processing first attempts to perform stemming using the standard or expanded inflectional stemmer, and then applies the stemmer guesser only to words that cannot be conventionally stemmed.

5.25.1.4 Part-of-Speech Tagging in Spanish

The following table shows the Spanish tag set. The tag names are accompanied by a brief description and one or more examples. If the example consists of more than one word, the word exemplifying the current tag is in **bold**. The tag set makes no distinction for gender.

Umbrella Tag	Complete Tag	Description	Examples
Adj	Adj	Invariant adjective	beige, mini
	Adj-Ord-Pl	Plural, spelled out ordinal number	primeros
	Adj-Ord-Sg	Singular, spelled out ordinal number	primer, primera, sexta
	Adj-Pl	Plural adjective	bonitos, nacionales
Adv	Adv	Adverb	siempre, directamente
	Adv-Deg	Adverbs that can modify adjectives	**muy** importante
	Adv-Int	Interrogative adverb	cuándo
	Adv-Rel	Adverbial relativizers	donde
Aux	Aux-be	The auxiliaries **ser** and **estar** ('be')	es, fui, estaba
	Aux-have	The auxiliary **haber** ('have')	han, hubo, hay
	Aux-Inf-be	Infinitive 'be' auxiliary	estar
	Aux-Inf-have	Infinitive form of **haber**	haber
	Aux-Inf-Pron-be	Infinitive be auxiliary with attached clitic	serme, estarlo
	Aux-Inf-Pron-have	Infinitive of **haber** with attached clitic	haberle, habérseme
Conj	Conj	Conjunction	si, porque
	Conj-como	The word **como**	como
	Conj-Coord	Coordinating conjunction	y, o
	Conj-que	The word **que**	que

Umbrella Tag	Complete Tag	Description	Examples
Det/Pron	Det/Pron-Quant-Pl	Plural quantifying determiner or pronoun	**unas** casas
	Det/Pron-Quant-Sg	Singular quantifying determiner or pronoun	poca
Det	Det-Dem-Pl	Plural demonstrative determiner	estas, esos
	Det-Dem-Sg	Singular demonstrative determiner	esta
	Det-Pl	Plural determiner	tus
	Det-Pre-Pl	Plural pre-determiner	**todas** las, **todos** los
	Det-Pre-Sg	Singular pre-determiner	**todo** el, **toda** la
	Det-Rel	Relative determiner	cual, cuyo
	Det-Sg	Singular determiner	mi
Interj	Interj	Interjection or onomatopoeia	ah
Nn	Nn	Noun, invariant for number	fénix
	Nn-Letter	Lowercase letter with or without a period or an uppercase without a period	a, h., M
	Nn-Net	URL or e-mail address	www.inxight.com, info@inxight.com
	Nn-Pl	Plural noun	gatos
	Nn-Sg	Singular noun	gato
Num	Num	Numeric expression, cardinal number	123, XIX, once, cuatrocientos
	Num-Ord	Ordinal number	1o., 2a
Part	Part-Neg	The negation particle **no**	no

Umbrella Tag	Complete Tag	Description	Examples
Prep	Prep	Preposition	en, con, por
	Prep-a	Preposition **a**	**a** casa
	Prep-de	Preposition **de**	la casa **de** María
	Prep-Det	Combination of preposition and determiner	fuera **del**, antes **del**
	Prep-Det-a	Combination **a** and determiner	al
	Prep-Det-de	Combination **de** and determiner	del
	Prep-para	Preposition **para**	**para** la casa
Pron	Pron	Pronoun	yo
	Pron-Clitic	Clitic pronoun (**acc.** or **dat.**)	le, me, os, nos
	Pron-Dem	Demonstrative pronoun	ésta, aquél
	Pron-Int	Interrogative pronoun	cuánto, cuál, quién
	Pron-Poss	Possessive pronoun	el **mío**, las **vuestras**
	Pron-Rel	Relative pronoun	lo cual, quien
	Pron-se	The reflexive pronoun	se
Prop	Prop	Proper noun or alphanumeric-punctuation combinations	Pablo U-587, Win2000
Punct	Punct	Other punctuation	' " / & { :;
	Punct-Comma	Comma	,
	Punct-Sent	Sentence punctuation	. ? !
V/Adj	V/Adj-PaPart-Pl	Plural past participle verb or adjective	hechas
	V/Adj-PaPart-Sg	Singular past participle verb or adjective	fundada

Umbrella Tag	Complete Tag	Description	Examples
V	V-Fin	Finite verb	tiene, pueda, dicte
	V-Impv	Imperative verb	dejad
	V-Impv-Pron	Imperative verb with attached clitic	déjame, sígueme
	V-Inf	Infinitive verb	evitar, tener, conducir
	V-Inf-Pron	Infinitive verb with attached clitic	hacerse, suprimirlas
	V-PrPart	Present participle verb	siendo, tocando
	V-PrPart-Pron	Present participle verb with attached clitic	haciéndoles, tomándolas

5.25.1.4.1 Unfound Words

Words not found in the tagger dictionary are passed to the Spanish tagger guesser where they are assigned a tag based on a set of rules about Spanish morphology and capitalization. The following set of tagging rules are part of this module.

Verb tags are assigned according to the verb conjugation patterns. Internet and e-mail addresses are tagged as `Nn-Net`.

Words beginning with a capital letter or a number followed by a capital letter are guessed as proper nouns. Combinations of alphabetic, numeric and optionally, punctuation characters are also guessed as proper nouns. Combinations of digits and punctuation are tagged as numbers. A series of punctuation marks is tagged as punctuation.

5.25.2 Extraction

This section describes the extraction-specific information for Spanish.

5.25.2.1 Spanish Subtypes

Spanish supports subtypes in the types `ORGANIZATION`, `REGION`, `SOCIAL_MEDIA`, and `URI`.

Related Topics

- Subtypes

5.25.2.2 Predefined Entity Types

This section describes the predefined entity types supported by the Spanish language module and examples of each. Click each link to jump to that subsection:ADDRESS1, COUNTRY, CURRENCY, DATE, DAY, GEO_AREA, GEO_FEATURE, HOLIDAY, LANGUAGE, LOCALITY, MEASURE, MONTH, NOUN_GROUP, ORGANIZATION, PEOPLE, PERCENT, PERSON, PHONE, PRODUCT, PROP_MISC, REGION, SOCIAL_MEDIA, TIME, TIME_PERIOD, TITLE, URI, and YEAR.

5.25.2.2.1 ADDRESS1

The format for `ADDRESS1` is based on typical postal address patterns found in Peninsular, Mexican and South American Spanish addresses.

An address must contain a street name and number. It may also include a city with state and/or country, a zip code or a neighborhood designator (used in Mexico).

- Avenida Cristobal Colón 5667
- Plaza de la Lealtad, 5 28014 Madrid
- Calle Castillo Chapultepec 47

  Colonia Chapultec

  62380 Cuernavaca, México
- Apartado Postal 20818, 28011 Madrid

5.25.2.2.2 COUNTRY

Names of countries and abbreviations for a limited set of countries. It also includes names of geopolitical entities for which conventional labels do not apply, such as disputed territories or territories that have not been internationally recognized . For example,

- Rusia
- Nicaragua
- Estado de Israel
- EE UU
- Kosovo
- Gibraltar
- Palestina

- Tibet

5.25.2.2.3 CURRENCY

Expressions denoting amounts of world currencies and expressions denoting ranges of these amounts:

- tres mil cuatrocientos veinte escudos
- 3 pesos chilenos
- $15
- ¥ 2500
- de 3 a 5 pesetas
- $15-30
- entre cinco mil y un millón de florines

5.25.2.2.4 DATE

Dates consisting of numbers, in several formats:

- 15-9-96
- 15.09.96
- 10/10/2001
- 2000-3-31
- 1980/05/02

Full dates must have at least a number and month:

- 31 de junio
- 1 de enero de 1555

5.25.2.2.5 DAY

Days of the week:

- lunes
- Miércoles
- viernes

Day ranges:

- lunes-viernes
- sábado/domingo

5.25.2.2.6 GEO_AREA

A geographical area larger than a city, including continents, groups of countries, and parts of continents:

- Alpes
- Norteamérica
- Centroamérica
- Caribe
- Cáucaso
- América Latina
- Europa del Este
- Medio Oriente
- Amazonía
- Balcanes

5.25.2.2.7 GEO_FEATURE

All place names that do not fall under `COUNTRY`, `GEO_AREA` , `LOCALITY`, or `REGION` :

- Cisjordania
- Mesopotamia
- Costa Azul
- Sierra Nevada
- Cabo Cañaveral
- Atlantis
- Valle de María
- Antillas
- Parque Nacional Nahuel Huapi
- Parque Nacional Galápagos
- Carretera Panamericana

5.25.2.2.8 HOLIDAY

Names of holidays and special days:

- Navidad
- Epifanía
- Semana Santa

- Nochevieja
- Sagrado Corazón
- Año Nuevo

5.25.2.2.9 LANGUAGE

Nouns referring to languages:

- el español
- el ruso
- el alemán
- el noruego
- el francés

5.25.2.2.10 LOCALITY

A city name:

- Madrid
- Tel Aviv
- Miami
- México
- Barcelona
- Roma

5.25.2.2.11 MEASURE

Any measurement, such as weight, volume or length:

- 3 km
- 9 grados
- 12 grados centígrados
- 75 kilos
- 348.000 hectáreas
- siete litros
- diez millas
- 646 toneladas
- 660 megavatios

Ranges of measurements and measurement expressions:

- de 28 a 30 grados
- 130 litros por metro cuadrado
- de 20.000 a 348.000 hectáreas

Ratios:

- cincuenta kilómetros por hora

5.25.2.2.12 MONTH

Month names, including abbreviations:

- octubre
- Jul

Month ranges with a hyphen or slash:

- julio-agosto
- julio/agosto

5.25.2.2.13 NOUN_GROUP

A Spanish simple noun phrase is a noun with optional pre- and postmodifiers:

- torneo femenino
- cascos azules
- decisión definitiva
- relaciones sino-norteamericanas
- verdadera reanudación
- préstamos inmobiliarios
- equilibrio presupuestario

5.25.2.2.14 ORGANIZATION

Commercial, governmental, educational, legal, and service agencies, including non-profit organizations, fine arts groups, and other associations and institutions, extracted as one of the following subtypes:

- `COMMERCIAL`–The name of commercial organizations, such as major companies or corporations. For example:
  - Texaco
  - Yamaha
  - agencia de noticias Xinhua

- FIAT
- Corporación del Cobre de Chile
- la empresa SOGEMIN
- Compañía de Energía de Ceara

Note:
Any extracted `ORGANIZATION/COMMERCIAL` entities can be parsed and standardized using the Data Quality Data Cleanse transform by mapping them to one of the `FIRM` input fields.

- `EDUCATIONAL`–The names of institutions focused primarily on education, for example:
  - Universidad Nacional Autónoma de México
  - Universidad Complutense de Madrid
  - Escuela Elemental Rafael De Jesús
- `OTHER`–Any other non-commercial organization including groupings of geopolotical entities that can function as political entities:
  - Consejo de Seguridad
  - UNESCO
  - Consejo Estatal Chino
  - Frente Zapatista de Liberación Nacional
  - Movimiento Revolucionario Túpac Amaru
  - Cámara de Comercio
  - Hospital Americano de Paris
  - Benelux
  - Comunidad Económica Europea
  - Unión Europea

5.25.2.2.15 PEOPLE

Names referring to identifiable groups of people based on country, ethnicity, region, or religion:

- los británicos
- los sirios
- los mexicanos
- los indígenas
- los brasileños
- los ecuatorianos

5.25.2.2.16 PERCENT

Percent expressions and expressions denoting ranges of percents:

- 60%
- 53.83%
- de 0,8 % a 16,44 %
- un cinco por ciento

5.25.2.2.17 PERSON

A person referred to by name. A variety of forms will be identified:

- Roberto
- Suzanne Prou
- Yitzhak Rabin
- PAPA JUAN PABLO II
- Juan Caballero Velásquez

Full or last name with a full or abbreviated title:

- Señor García
- Sr. Sanchez-Farrés
- Sra. María José de la Garza

5.25.2.2.18 PHONE

Fax and telephone numbers used in Spanish-speaking countries:

- (331) 40 41 45 69
- Fax: (331) 40 41 46 95
- Tel: 34 91 33782 00
- 1-800-111-2222
- Tel: 91 111 11 11
- +34 111 222222
- 111 222 3333 ext 1111
- Fax: 111-2222
- 111-2222

5.25.2.2.19 PRODUCT

A product name, optionally preceeded by a company name:

- Boeing 757
- Marlboro
- Windows
- Compaq 3-5/8

5.25.2.2.20 PROP_MISC

Any proper noun phrase that does not belong to one of the entity types specified by the other entities:

- Prestige
- Wye Plantation
- Guatemala-Elecciones
- Zimbabue/N.8

5.25.2.2.21 REGION

Different regions are extracted as one of the following subtypes:

- `MAJOR-` A state or an administrative division (such as a provinces or autonomous region) of a country:
  - País Vasco
  - Canarias
  - Chiapas
  - provincia de Córdoba
  - Cauca
  - California
  - Minas Gerais
- `MINOR-` Names of counties, prefectures, districts, or analogous geographical divisions or governmental units:
  - Martinica
  - Guadalupe

5.25.2.2.22 SOCIAL_MEDIA

Entity type for extracting entities from social media feeds. The handles (also known as `ID`) and topics are extracted as one of the following subtypes:

Note:
The `SOCIAL_MEDIA` entity type supports only Twitter feeds.

- `ID_TWITTER`–Twitter handles or `ID`s starting with "`@`", for example:
  - `@IsabelNevado`
  - `@Ramón_Sanchez`
  - `@SCNblogs`
  - `@sapnoticiasbr`
  - `@sapnews`
  - `@SAP_MICROSOFT`
- `TOPIC_TWITTER`–Twitter topics starting with "`#`", for example:
  - `#EnElFuturo`
  - `#La_Colonia`
  - `#SAP`
  - `#Mobility`
  - `#SAPPRESS`
  - `#SAP_projects`

5.25.2.2.23 TIME

Clock times and time expressions:

- 13:45
- 1:45 de la tarde
- la 1.45 de la tarde
- las 2:30 horas
- 12H45
- las 08H00
- 07h GMT
- 05H00 GMT
- LAS 12H00 GMT

Clock time expressions in words:

- las cinco y cuarto de la tarde
- las diez de la mañana

5.25.2.2.24 TIME_PERIOD

Measures of time duration and expressions denoting ranges of measures:

- doce horas
- 15 minutos
- cuatro décadas
- 20 meses
- cinco siglos

5.25.2.2.25 TITLE

An individual specified by position or title, without an accompanying name:

- Rey
- Subcomandante
- Secretario de Estado

5.25.2.2.26 URI

An address on the internet, extracted as one of the following subtypes:

- `EMAIL`– Email addresses, including Lotus Notes email addresses , for example:
  - ana.sanchez@comercio.es
  - Dupont/BOBJ@CMP
  - CTarin/Inxight@CMP
- `URL`– Internet addresses, for example:
  - elpais.es
  - www.elpais.es

5.25.2.2.27 YEAR

Year identifiers, in full or abbreviated form:

- 1982
- 444 aC

- '68
- '50 y '60

Decade or century identifiers:

- los años sesenta
- la década de los noventa
- los setenta
- siglo XX

Year ranges:

- 1979-90
- entre 1989 y 1991

5.26 Swedish Language Reference

This chapter describes the behavior of the Swedish language module.

5.26.1 Linguistic Processing

This section describes the language-specific information on the linguistic processing of Swedish texts, including word segmentation and stemming.

5.26.1.1 Character Encodings for Swedish

- iso_8859_1
- cp_1252
- utf_8, utf_16, ucs_4

5.26.1.2 Word Segmentation in Swedish

The Swedish segmenter follows all of the general segmentation rules in the white space languages. The Swedish segmenter has the following language-specific behavior.

The Swedish segmenter does not split plurals and possessives spelled with **s** or **'s**. Hyphens are not separated from compound parts written with a hyphen. Numeric and punctuation combinations are kept together.

Text	Segmented
Eriks	Eriks
metall- och kemikoncern	metall-
	och
	kemikoncern
456:-	456:-

Related Topics

- Word Segmentation
- White Space Languages

5.26.1.3 Stemming in Swedish

This section describes the standard stemmer, the expanded inflectional stemmer, and the compound stemmer used for stemming in Swedish.

5.26.1.3.1 Standard Stemmer

The Swedish stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Baseform	Examples
Noun	Indefinite singular	hunden, hundar, hundarna -> hund
Verb	Infinitive	springer, sprang, sprungit -> springa
Adjective	Base form	vackra, vackert -> vacker
Adverb	Base form or source form	snabbt -> snabbt

5.26.1.3.2 Expanded Inflectional Stemmer

The expanded inflectional stemmer allows certain non-standard word forms–for example, capitalization errors–as well as standard forms, and thus can be used to process informal or imperfect text (such as email, online documents, or queries). See Expanded Inflectional Stemming for the general behavior. The following lists the specifics for Swedish.

The expanded version accepts lower case letters in addition to capital letters for words that are usually capitalized.

Example	Output
Erik	Erik
erik	Erik

Typewriter Forms of Accented Letters

The expanded version accepts typewriter conventions for accented letters. That is, **å** is recognized when written as **aa**, **ä** when written as **ae**, and **ö** when written as **oe**.

Example	Output
blaa	blå
blå	blå

Deaccented Characters

The expanded version also allows deaccented characters in place of accented ones.

Example	Output
bla	blå
blå	blå

Hyphenation

To aid software that handles line-breaking hyphens by deleting them and concatenating the two parts of the broken word, hyphens in non-numeric expressions are optional in the expanded version, so that words that are truly hyphenated will still be recognized.

Example	Output
uland	u-land
u-land	u-land

5.26.1.3.3 Compound Stemmer

The compound stemmer performs standard inflectional stemming along with the stemming of productive compounds. Specifically, the compound stemmer breaks up compounds and stems the parts. Compounds that are productive are not stored in a dictionary, thereby enabling various combinations of words.

Note:
The standard stemmer does not perform compound analysis.

The following examples illustrate the operation of the compound stemmer:

Noun-noun

Example	Output
aftonbladet	afton | bladet
berglandskap	berg | landskap

Noun ninus final -a/e + noun

Example	Output
flickskola	flick | skola (from flicka and skola)
kistnyckel	kist | nyckel (from kista and nyckel)

Noun+s+noun

Example	Output
anfallsspel	anfall | spel
stavningsreform	stavning | reform

5.26.1.4 Part-of-Speech Tagging in Swedish

The following table shows the Swedish tag set available for using in defining custom entities. The tag names are accompanied by a brief description and one or more examples. If the example consists of more than one word, the word exemplifying the current tag is in **bold**. The tag set makes no distinction for gender.

Umbrella Tag	Complete Tag	Description	Examples
`Abbr`	`Abbr`	Abbreviation	S.
`Adj`	`Adj-Comp`	Comparative adjective	äldre
	`Adj-Def-Pl`	Definite and/or plural adjective	svenska
	`Adj-Indef-Sg`	Indefinite singular adjective	grov
	`Adj-Sup`	Superlative adjective	viktigast
	`Adj-Sup-Def`	Definite superlative adjective	fullaste

Umbrella Tag	Complete Tag	Description	Examples
Adv	Adv	Adverb	redan
Cmpd	Cmpd-Part	Compound part	**plats-** och släktnamnen
Conj	Conj	Conjuction	att
	Conj-Coord	Coordinating conjunction	och, eller
Det/Pron	Det/Pron-Pl	Plural determiner or pronoun	dessa
	Det/Pron-Pl-Gen	Plural genitive determiner or pronoun	andras
	Det/Pron-Sg	Singular determiner or pronoun	något, denna
	Det/Pron-Sg-Gen	Singular genitive determiner or pronoun	dennes
Det	Det-Def-Pl	Definite plural determiner	**de** partierna
	Det-Def-Sg	Definite singular determiner	**den** ekonom
	Det-Indef-Sg	Indefinite singular determiner	en, ett
Interj	Interj	Interjection	ja

Umbrella Tag	Complete Tag	Description	Examples
Nn	Nn-Def-Pl	Definite plural noun	verken
	Nn-Def-Pl-Gen	Genitive definite plural noun	professionernas
	Nn-Def-Sg	Definite singular noun	historikern
	Nn-Def-Sg-Gen	Genitive definite singular noun	dagens
	Nn-Indef-Pl	Indefinite plural noun	strukturer
	Nn-Indef-Pl-Gen	Genitive indefinite plural noun	rikens
	Nn-Indef-Sg	Indefinite singular noun	dag
	Nn-Indef-Sg-Gen	Genitive indefinite singular noun	institutions
	Nn-Indef-SP	Indefinite singular or plural noun	universitet
	Nn-Indef-SP-Gen	Genitive indefinite singular or plural noun	års
	Nn-Letter	Lowercase and uppercase letters	b, N
	Nn-Net	URL and e-mail address	www.inxight.com info@inxight.com
	Nn-Def-Pl	Definite plural noun	verken
Num	Num	Cardinal number, in digits or spelled-out plural number	5,4 or 300 hundra
	Num-Sg	Spelled-out number "one"	en
Ord	Ord	ordinal number (digits or words)	tredje

Umbrella Tag	Complete Tag	Description	Examples
Prep	Prep	Preposition	kring
	Prep-av	Preposition **av**	av
	Prep-foer	Preposition **för**	för
	Prep-om	Preposition **om**	om
	Prep-paa	Preposition **på**	på
Pron	Pron-Acc	Accusative pronoun	en
	Pron-Gen	Genitive pronoun	ens
	Pron-Nom	Nominative pronoun	man
	Pron-Pers-Acc	Accusative personal pronoun	sig
	Pron-Pers-Gen	Genitive personal pronoun	dess, deras
	Pron-Pers-Nom	Nominative personal pronoun	du
	Pron-Poss-Pl	Plural possessive pronoun	mina
	Pron-Poss-Sg	Singular possessive pronoun	vår, ert
Prop	Prop	Proper name	Europa, Margareta
	Prop-Gen	Genitive proper name	Eriks
Punct	Punct	Punctuation	/ -
	Punct-Comma	Comma	,
	Punct-Paren	Bracketing punctuation	( ) [ ]
	Punct-Quote	Quotation punctuation	" ' ' « »
	Punct-Sent	Sentence terminating punctuation	. ? !
Quant	Quant	Quantifier	alla, många

Umbrella Tag	Complete Tag	Description	Examples
V	V-Impv	Imperative verb	tänk
	V-Impv-SForm	Imperative verb, S-Form	minns
	V-Inf	Infinitive verb	skilja
	V-Inf-SForm	Infinitive verb, S-Form	tänkas
	V-PaPart	Past participle verb	tecknat, handlat
	V-Past	Past tense verb	slog
	V-Past-SForm	Past tense verb, S-Form	kysstes
	V-Pres	Present tense verb	varnar
	V-Pres-SForm	Present tense verb, S-Form	sägs
	V-PrPart	Present participle verb	mötande

5.26.1.5 Grouping in Swedish

Swedish noun phrases contain one or more nouns with optional modifiers. Noun modifiers can be the genitive form of nouns, a series of adjectives (possibly with adverbs), and participial phrases. For example:

- svenska småföretag
- skiftande arbetstider
- årligen återkommande attraktion
- Göteborgs stadsteater

Coordination of modifiers is allowed, as well as compound parts, as in:

- metall- och kemikoncern
- socialdemokratiska ledare och regeringschefer

The prepositions **för**, **om**, **av**, and **på** are allowed in noun phrases. For example:

- knepig fråga för regeringen

- debatt om rasism
- behandling av naturliga språk
- synpunkt på de övriga skäl

5.26.2 Extraction

This section describes the extraction-specific information for Swedish.

5.26.2.1 Predefined Entity Types

This section describes the predefined entity types supported by the Swedish language module and examples of each.

5.26.2.1.1 NOUN_GROUP

Swedish noun phrases contain one or more nouns with optional modifiers. Noun modifiers can be the genitive form of nouns, a series of adjectives (possibly with adverbs), and participial phrases. For example:

- svenska småföretag
- skiftande arbetstider
- årligen återkommande attraktion
- Göteborgs stadsteater

Coordination of modifiers is allowed, as well as compound parts, as in:

- metall- och kemikoncern
- socialdemokratiska ledare och regeringschefer

5.27 Thai Language Reference

This chapter describes the behavior of the Thai language module.

5.27.1 Linguistic Processing

This section describes the language-specific information on the processing of Thai texts, including word segmentation and stemming.

5.27.1.1 Character Encodings for Thai

- tis_620
- utf_8, utf_16, ucs_4

5.27.1.2 Word Segmentation in Thai

The Thai segmenter follows all of the general segmentation rules in the non-white space languages. Thai has the following language-specific behavior.

Suffixes and prefixes remain attached to their content words, as shown below:

Text	Segmented
ผู้สอม	ผู้สอม
โดยตั้งใจ	โดยตั้งใจ

Related Topics

- Word Segmentation

5.27.1.3 Stemming in Thai

Since Thai words are not inflected, the stems of all words are identical to their source forms. This is true of the open class words listed in the following table as well as the closed class words.

Category	Baseform	Examples
Noun	Source form	ภาพถ่าย -> ภาพถ่าย
Verb	Source form	กลั่น -> กลั่น
Adjective	Source form	คอนกรีต -> คอนกรีต
Adverb	Source form	กรอด -> กรอด

5.27.2 Extraction

Note:
Thai is a basic-level supported language module, which means it supports extraction by using dictionaries or extraction rules only.

5.28 Turkish Language Reference

This chapter describes the behavior of the Turkish language module.

5.28.1 Linguistic Processing

This section describes the language-specific information on the processing of Turkish texts, including word segmentation and stemming.

5.28.1.1 Character Encodings for Turkish

- iso_8859-9
- utf_8, utf_16, ucs_4

5.28.1.2 Word Segmentation in Turkish

The Turkish segmenter follows all of the general segmentation rules in the white space languages.

Related Topics

- Word Segmentation

5.28.1.3 Stemming in Turkish

The Turkish stemmer follows the general stemming rules, as described in Stemming. In brief, the major word classes, also known as the open classes, stem to their baseforms. This is shown in the table below. Closed class words like determiners and pronouns may also be stemmed.

Category	Examples
Noun	kitaplar -> kitap, anahtarýný -> anahtar, denizi -> deniz
Verb	açabilir -> aç, gideceðim -> git, boyamaya -> boya
Adjective	küçük -> küçük, dikkatli -> dikkat, güzel -> güzel
Adverb	sessiz -> ses, gergin -> ger, çabukça -> çabuk

5.28.2 Extraction

Note:
Turkish is a basic-level supported language module, which means it supports extraction by using dictionaries or extraction rules only.

Voice of the Customer Content

Voice of the customer content includes a set of entity types and rules that address requirements for extracting customer sentiments and requests. You can use this content to retrieve specific information about your customers' needs and perceptions when processing and analyzing text.

Voice of the customer content is included in and supports these language modules:

- English
- French
- German
- Spanish

Extraction involves complex linguistic analysis and pattern matching that includes processing parts of speech, syntactic patterns, negation, and so on, to identify the patterns to be extracted.

The extraction output includes the identified patterns and information about each extraction, including the type of information extracted (either sentiment or request).

Voice of the customer content includes the following rule sets and dictionaries:

Rule Set Description	Compiled (.fsm) and Source (.rul) Files

• Sentiment
Extracts information about sentiments and problems, including the strength of the sentiment, such as strong or weak
	`english-tf-voc-sentiment.fsm`
	`english-tf-voc-sentiment.rul`
	`french-tf-voc-sentiment.fsm`
	`french-tf-voc-sentiment.rul`
	`german-tf-voc-sentiment.fsm`
	`german-tf-voc-sentiment.rul`
	`spanish-tf-voc-sentiment.fsm`
	`spanish-tf-voc-sentiment.rul`

Rule Set Description	Compiled (.fsm) and Source (.rul) Files
• Emoticon Extracts information about sentiments expressed by emoticons, which generally convey the user's feelings towards the whole utterance and not a particular topic within it	`english-tf-voc-emoticon.fsm` `english-tf-voc-emoticon.rul`

• Request
Extracts general requests made by the customer, including requests to be contacted
	`english-tf-voc-request.fsm`
	`english-tf-voc-request.rul`
	`french-tf-voc-request.fsm`
	`french-tf-voc-request.rul`
	`german-tf-voc-request.fsm`
	`german-tf-voc-request.rul`
	`spanish-tf-voc-request.fsm`
	`spanish-tf-voc-request.rul`

When analyzing and extracting data, the rule sets consider a statement that expresses a customer sentiment or a request as a disposition. Dispositions are further divided into stances and topics:

- For sentiment dispositions, the stance represents the category or type of sentiment, for example, a strong positive sentiment or a strong negative sentiment.
- For a request disposition, the stance represents the type of request: general or contact.
- Emoticon rule sets extract only stances, i.e. sentiments, without identifying topics.
- A topic represents what the sentiment or the request is about.

Each extracted disposition typically includes:

- At least one stance
- An optional topic; multiple topics are allowed within one disposition.

Dictionary Description	Compiled (.nc) and Source (.xml) Files

• Profanity
Extracts pejorative words and phrases
	`english-tf-voc-AmbigProfanity.nc`
	`english-tf-voc-AmbigProfanity.xml`
	`english-tf-voc-UnambigProfanity.nc`
	`english-tf-voc-UnambigProfanity.xml`

For details about using voice of the customer content to enhance extraction, refer to the *SAP BusinessObjects Data Services Text Data Processing Extraction Customization Guide*.

Related Topics

- Extracting Sentiments
- Extracting Emoticons
- Extracting Requests
- Extracting Profanities

6.1 Extracting Sentiments

Sentiment rules are designed to extract information about someone's feelings about something. The rules extract patterns that express customer feelings about concepts, places, actions, items, and so forth; for example, a product, company, service, or person.

The rules categorize extracted sentiments into the following types of stances:

Type of Sentiment Stance	Description
Strong positive sentiment	A strong positive opinion, such as "great" or "excellent"
Weak positive sentiment	A weak positive opinion, such as "nice" or "fine"
Neutral sentiment	An opinion that is not positive or negative, such as "OK" or "acceptable"
Weak negative sentiment	A weak negative opinion, such as "bad" or "dislike"
Strong negative sentiment	A strong negative opinion, such as "hate" or "terrible"
Minor problem	An opinion describing an impediment the customer can work around, such as "useless" or "faulty"

Type of Sentiment Stance	Description
Major problem	An opinion describing an impediment the customer cannot work around, such as "broke down" or "not work"

In addition to a stance, an extracted sentiment optionally includes one topic–the subject of the sentiment. The topic answers the question "What is it that the customer is expressing their feelings about?"

6.1.1 English: Sentiment Extraction Examples

Strong Positive Sentiment

- I totally recommend this car for everyone.

  `[Sentiment]`I `[StrongPositiveSentiment]`totally recommend`[/StrongPositiveSentiment] [Topic]`this car`[/Topic]` for everyone.`[/Sentiment]`

Weak Positive Sentiment

- The screen is nice with a good picture quality.

  `[Sentiment][Topic]`The screen`[/Topic]` is `[WeakPositiveSentiment]`nice`[/WeakPositiveSentiment]` with a `[WeakPositiveSentiment]`good`[/WeakPositiveSentiment] [Topic]`picture quality`[/Topic].[/Sentiment]`

Neutral Sentiment

- I don't love your software.

  `[Sentiment]`I do`[NeutralSentiment]`n't love`[/NeutralSentiment] [Topic]`your software`[/Topic].[/Sentiment]`

Weak Negative Sentiment

- I was disappointed that the coffee was cold.

  `[Sentiment]`I was `[WeakNegativeSentiment]`disappointed`[/WeakNegativeSentiment] [Topic]`that the coffee was cold`[/Topic][/Sentiment].[/Sentiment]`

Strong Negative Sentiment

- I am very dissatisfied with your service.

  `[Sentiment]`I am `[StrongNegativeSentiment]`very dissatisfied`[/StrongNegativeSentiment]` with `[Topic]`your service`[/Topic].[/Sentiment]`

Minor Problem

- Most flights are overbooked.

  [Sentiment][Topic]Most flights[/Topic] are [MinorProblem]overbooked[/MinorProblem].[/Sentiment]

Major Problem

- Contrary to what the package indicates, it doesn't work.

  [Sentiment]Contrary to what the package indicates, it does[MajorProblem]n't work[/MajorProblem].[/Sentiment]

6.1.2 French: Sentiment Extraction Examples

Strong Positive Sentiment

- Les repas y sont excellents.

  [Sentiment][Topic]Les repas[/Topic] y sont [StrongPositiveSentiment]excellents[/StrongPositiveSentiment][/Sentiment].

Weak Positive Sentiment

- Je suis satisfait du confort intérieur.

  Je suis [Sentiment][WeakPositiveSentiment]satisfait[/WeakPositiveSentiment] du [Topic]confort intérieur[/Topic][/Sentiment].

Neutral Sentiment

- Une voiture relativement acceptable.

  [Sentiment][Topic]Une voiture[/Topic] [NeutralSentiment]relativement acceptable[/NeutralSentiment][/Sentiment].

Weak Negative Sentiment

- Un serveur assez impoli.

  [Sentiment][Topic]Un serveur[/Topic] [WeakNegativeSentiment]assez impoli[/WeakNegativeSentiment][/Sentiment].

Strong Negative Sentiment

- Le tirage photo me déçoit beaucoup par sa mauvaise qualité.

  [Sentiment][Topic]Le tirage photo[/Topic] me [StrongNegativeSentiment]déçoit beaucoup[/StrongNegativeSentiment][/Sentiment] par sa mauvaise qualité.

Minor Problem

- Le problème se situe dans le moteur.

  Le [Sentiment][MinorProblem]problème[/MinorProblem] se situe dans le moteur.[/Sentiment]

Major Problem

- Contrairement à ce que l'emballage indique, ça ne fonctionne pas.

  Contrairement à ce que l'emballage indique, [Sentiment]ça ne [MajorProblem]fonctionne pas[/MajorProblem][/Sentiment].

6.1.3 German: Sentiment Extraction Examples

Strong Positive Sentiment

- Der Fernseher ist klasse.

  [Sentiment][Topic] Der Fernseher [/Topic] ist [StrongPositiveSentiment] klasse [/StrongPositiveSentiment][/Sentiment].

Weak Positive Sentiment

- Ich mag das Radio.

  [Sentiment] Ich [WeakPositiveSentiment] mag [/WeakPositiveSentiment][Topic] das Radio[/Topic][/Sentiment].

Neutral Sentiment

- Ich finde Ihre Produkte etwas mittelmäßig.

  Ich finde [Sentiment][Topic] Ihre Produkte[/Topic][NeutralSentiment] etwas mittelmäßig [/NeutralSentiment][/Sentiment].

Weak Negative Sentiment

- Der Hauptbahnhof ist nicht schön.

  [Sentiment][Topic] Der Hauptbahnhof [/Topic] ist [WeakNegativeSentiment] nicht schön[/WeakNegativeSentiment][/Sentiment].

Strong Negative Sentiment

- Der Service war furchtbar.

  [Sentiment][Topic] Der Service [/Topic] war [StrongNegativeSentiment] furchtbar [/StrongNegativeSentiment][/Sentiment].

Minor Problem

- Es sieht so aus, als ob es meinem Computer schadet.

  Es sieht so aus, als ob es [Sentiment][Topic] meinem Computer [/Topic][MinorProblem] schadet [/MinorProblem][/Sentiment].

Major Problem

- Die Installierung hat meinen Computer kaputt gemacht!.

  [Sentiment] Die Installierung hat [Topic] meinen Computer [/Topic][MajorProblem] kaputt gemacht [MajorProblem][/Sentiment].

6.1.4 Spanish: Sentiment Extraction Examples

Strong Positive Sentiment

- Absolutamente adoro este álbum.

  [Sentiment][StrongPositiveSentiment]Absolutamente adoro[/StrongPositiveSentiment] [Topic]este álbum[/Topic][/Sentiment].

Weak Positive Sentiment

- Me gusta este grupo.

  [Sentiment][WeakPositiveSentiment]Me gusta[/WeakPositiveSentiment] [Topic]este grupo[/Topic][/Sentiment].

Neutral Sentiment

- No es una maravilla.

  [Sentiment]No es una [NeutralSentiment]maravilla[/NeutralSentiment][/Sentiment].

Weak Negative Sentiment

- Es una mala tienda.

  [Sentiment]Es una [WeakNegativeSentiment]mala[/WeakNegativeSentiment] [Topic]tienda[/Topic][/Sentiment].

Strong Negative Sentiment

- Odio este televisor.

  [Sentiment][StrongNegativeSentiment]Odio[/StrongNegativeSentiment] [Topic]este televisor[/Topic][/Sentiment].

Minor Problem

- Tengo problemas con el sonido.
- `[Sentiment]`Tengo `[MinorProblem]`problemas`[/MinorProblem]` con `[Topic]`el sonido`[/Topic][/Sentiment].`

Major Problem

- El archivo es corrupto.

  `[Sentiment][Topic]`El archivo`[/Topic]` es `[MajorProblem]`corrupto`[/MajorProblem][/Sentiment].`

6.2 Extracting Emoticons

Emoticon rules are designed to extract infomation about someone's feelings about the whole sentence or situation. Unlike sentiment rules, emoticon rules do not specify a particular topic of the stance.

The rules categorize extracted emoticons into the following types of stances:

Type of Emoticon	Description
Weak positive emoticon	Extracts emoticons conveying weak positive sentiment, e.g. :-), ;-), ;), (:
Strong positive emoticon	Extracts emoticons conveying strong positive sentiment, e.g. :-D, :-))), (((:
Weak negative emoticon	Extracts emoticons conveying weak negative sentiment, e.g. :-(, :(, ;-\
Strong negative emoticon	Extracts emoticons conveying strong negative sentiment, e.g. :-(((, :((, :'-(

6.2.1 English: Emoticon Extraction Examples

Weak Positive Emoticon

- Loving my new BlackBerry! :-) No iPhone needed over here.

  `[Emoticon]`Loving my new BlackBerry! `[WeakPositiveEmoticon]`:-)`[/WeakPositiveEmoticon]` No iPhone needed over here. `[/Emoticon]`

Strong Positive Emoticon

- The show was hilarious :-D

  `[Emoticon]` The show was hilarious `[StrongPositiveEmoticon]`:-D`[/StrongPositiveEmoticon][/Emoticon]`

Weak Negative Emoticon

- I hate this phone i'm using :-(

  `[Emoticon]` I hate this phone i'm using `[WeakNegativeEmoticon]`:-(`[/WeakNegativeEmoticon][/Emoticon]`

Strong Negative Emoticon

- The Dow Jones fell 200 points :-(((

  `[Emoticon]` The Dow Jones fell 200 points `[StrongNegativeEmoticon]` :-(((`[/StrongNegativeEmoticon][/Emoticon]`

6.3 Extracting Requests

Request rules are designed to extract information about a customer's wish for a change or enhancement. The rules extract patterns that express customer requests to be contacted or for new or added functionality in an item such as a product, company, service, or person.

Requests are categorized into the following stances:

Type of Request Stance	Description
General request	A request for an enhancement or for something new, such as "please add", "please create", or "would like"
Contact request	A request for direct or immediate contact, such as "please contact me" or "call me"

An extracted request contains the following:

- One request stance (`ContactRequest` or `GeneralRequest`)
- Optionally, one topic–what the the request is about. The topic answers the question "What does the customer want?"
- Optionally, contact information–for phone or fax numbers, addresses, email addresses, web site addresses

6.3.1 English: Request Extraction Examples

General Requests

- Improve the software UI.

  `[Request][GeneralRequest]`Improve`[/GeneralRequest] [Topic]`the software UI`[/Topic].[/Request]`

- An additional switch would be great to have on this vacuum cleaner.

  `[Request][Topic]`An additional switch`[/Topic] [GeneralRequest]`would be great`[/GeneralRequest]` to have on this vacuum cleaner.`[/Request]`

Contact Requests

- I would like to be contacted by your customer support service at 617-555-5555.

  `[Request]`I `[ContactRequest]`would like to be contacted`[/ContactRequest]` by `[Topic]`your customer support service`[/Topic]` at `[ContactInfo]`617-555-5555`[/ContactInfo].[/Request]`

- I would like to receive the January catalog.

  `[Request]`I `[ContactRequest]`would like to receive`[/ContactRequest] [Topic]`the January catalog`[/Topic].[/Request]`

6.3.2 French: Request Extraction Examples

General Requests

- Le conso aimerait savoir s'il peut avoir une extension de garantie.

  `[Request]`Le conso `[GeneralRequest]`aimerait`[/GeneralRequest] [Topic]`savoir s'il peut avoir une extension de garantie`[/Topic][/Request].`

- J'aurais aimé trouver plus de fonctions.

[Request]J'[GeneralRequest]aurais aimé[/GeneralRequest] [Topic]trouver plus de fonctions[/Topic].[/Request]

Contact Requests

- Il souhaitait des informations sur les nouveaux produits.

  [Request]Il [ContactRequest]souhaitait[/ContactRequest] [Topic]des informations[/Topic] sur les nouveaux produits.[/Request]

- Il demande des renseignements sur la garantie constructeur.

  [Request][ContactRequest]Il demande[/ContactRequest] [Topic]des renseignements[/Topic] sur la garantie constructeur.[/Request]

6.3.3 German: Request Extraction Examples

General Requests

- Lizenzmodel deutlich vereinfachen!

  [Request] [Topic] Lizenzmodel [/Topic] [GeneralRequest] deutlich vereinfachen [/GeneralRequest] ! [/Request]

Contact Requests

- Rufen Sie mich unter der Nummer 555-1212 an.

  [Request] [ContactRequest] Rufen [/ContactRequest] Sie mich unter der Nummer [ContactInfo] 555-1212 [/ContactInfo] an. [/Request]

6.3.4 Spanish: Request Extraction Examples

General Requests

- Podría hacer un otro color en vez de blanco?

  [Request]Podría [GeneralRequest]hacer[/GeneralRequest] [Topic]un otro color en vez de blanco[/Topic]?[/Request]

Contact Requests

- Quiero contactarme contigo.

  [Request][ContactRequest]Quiero contactarme contigo[/ContactRequest][/Request].

6.4 Extracting Profanities

The aim of the Profanity dictionary is to define a set of pejorative vocabulary.

Profanities are categorized into two classes:

Type of Dictionary	Description
Ambiguous profanity	Extracts words and phrases that are pejorative only in certain contexts
Unambiguous profanity	Extracts words and phrases that are always pejorative, regardless of the context

6.4.1 English: Profanity Extraction Examples

Ambiguous Profanity

- Those hooligans threw toilet paper on my lawn.

  Those `[AMBIGUOUS_PROFANITY]`hooligans`[/AMBIGUOUS_PROFANITY]` threw toilet paper on my lawn.

Unambiguous Profanity

- I cannot express how angry I am with this asshole.

  I cannot express how angry I am with this `[UNAMBIGUOUS_PROFANITY]`asshole`[/UNAMBIGUOUS_PROFANITY]`.

Enterprise Content

The specialized enterprise content includes rules that address domain-specific extraction requirements for an enterprise. You can use this enterprise content to extract these specific types of information when processing and analyzing text:

Rule Set Description	Compiled (.fsm) and Source (.rul) Files
• Membership Information Extracts information about a person's affiliations	`english-tf-ent-Member.fsm` `english-tf-ent-Member.rul`
• Management Changes Extracts information about management changes	`english-tf-ent-ManagementChanges.fsm` `english-tf-ent-ManagementChanges.rul`
• Product Releases Extracts information about product releases	`english-tf-ent-ProductRelease.fsm` `english-tf-ent-ProductRelease.rul`
• Mergers and Acquisitions Extracts information about mergers and acquisitions	`english-tf-ent-Merger.fsm` `english-tf-ent-Merger.rul`
• Organizational Information Extracts information about an organization, such as founder, location, or contact information	`english-tf-ent-OrganizationInfo.fsm` `english-tf-ent-OrganizationInfo.rul`

Note:
The enterprise content is included in and supports the English language module only.

For details about using enterprise content to enhance extraction rules, refer to the *SAP BusinessObjects Data Services Text Data Processing Extraction Customization Guide*.

Related Topics

- Extracting Membership Information
- Extracting Management Change Events
- Extracting Product Release Events
- Extracting Merger Information
- Extracting Organizational Information

7.1 Extracting Membership Information

The member rules are designed to extract personal membership information about an individual, and the position held within the organization.

The following table describes the rules for extracting membership information:

Rule (though labeled only as "Member")	Description	Example
`Member_OrgPerPos`	Extracts patterns that follow the general form "Organization, Person, Position"	**Southern Community Financial Corporation** announced the appointment of **Richard M. Cobb, Executive Vice President, Chief Operating Officer and Chief Financial Officer**

Rule (though labeled only as "Member")	Description	Example
`Member_OrgPosPer`	Extracts patterns that follow the general form "Organization, Position, Person"	"It is a typical tight stock situation," said **Smith Barney analyst Walter Spilka**.
`Member_OrgPosPer_Pos Guess`		Baptist minister and **Salvation Army volunteer Ralph Balley** says the favorite song among inmates is the hymn "Amazing Grace" about a lost soul who found redemption.
`Member_OrgPosPer_Per Guess`		**CanWest Global Communications Corp Chief Executive Officer Izzy Asper** said on Thursday he expects the broadcast company will see a "substantial advance" in profits, dividends and developments in 1997.
`Member_PerOrgPos`	Extracts patterns that follow the general form "Person, Organization, Position"	**Larry Wachtel**, a **Prudential Securities market analyst**
`Member_PerOrgMember`		**Yael Dayan**, a **Labour Party member**
`Member_PerMemberOrgBoD`		**Willy Kiekens**, a member of the **IMF**'s **board of directors**
`Member_PerPosOrg`	Extracts patterns that follow the general form "Person, Position, Organization"	**Harry Reid, president** of **True North Communications International**
`Member_PerMemberOrg`		**Willy Kiekens**, a **member** of the **IMF**

7.2 Extracting Management Change Events

The management changes rules are designed to extract information related to a change in an individual's title and company, including any information about the previous or future title holder.

The management changes rules file includes these two groups of rules:

- `HireEvent`–These rules extract patterns related to the start of employment
- `ResignEvent`–These rules extract patterns related to the end of employment

Table 7-3: Hire Event Rules

Rule (though labeled only as "HireEvent" and "Resign Event")	**Description**	**Example**
`HireEvent_OrgActPerPos`	Extracts patterns that follow the general form "Organization, Action, Person, Position"	**Big Bear Networks**, delivering intelligence into the Optical/Electrical interface for enterprise, metro and carrier networks, today **named Amit Jain** as its **president and chief executive officer**.
`HireEvent_OrgPerActPos`	Extracts patterns that follow the general form "Organization, Person, Action, Position"	**Specialty Disease Management Services, Inc.**, a leading provider of disease management services, announced today that **Brian Vervynck** has **joined** the firm as **vice president of sales**.
`HireEvent_OrgPerActBoD`		**nLayers**, the leader in real-time discovery and resource optimization, announced today that **Frank Moss, Ph.D.**, has **joined** its **Strategic Advisory Board**.
`HireEvent_PerActPosOrg`	Extracts patterns that follow the general form "Person, Action, Position, Organization"	**Jeff McLean** has been **named president** of **CooperVision**'s U.S. operations.

Rule (though labeled only as "HireEvent" and "Resign Event")	**Description**	**Example**
`HireEvent_PerActOrgPos`	Extracts patterns that follow the general form "Person, Action, Organization, Position"	**Brad Jones joins CooperVision**'s management team as **vice president of U.S. sales**.
`HireEvent_PerActOrgBoD`		**Willy Kiekens** was **named** to the **IMF**'s **board of directors**.

Table 7-4: Resign Event Rules

Rule	**Description**	**Example**
`ResignEvent_OrgPerActPos`	Extracts patterns that follow the general form "Organization, Person, Action, Position"	**Suburban Propane Partners L.P.** said Thursday that **Salvatore Quadrino resigned** as **president** to pursue other business opportunities.
`ResignEvent_OrgPerPosAct`	Extracts patterns that follow the general form "Organization, Person, Position, Action"	**Southern Community Financial Corporation** (Nasdaq: SCMF; SCMFO) announced today that on February 5, 2005, **Richard M. Cobb**, **Executive Vice President, Chief Operating Officer and Chief Financial Officer of Southern Community Financial Corporation** , announced his **resig nation** from the Company effective today.
`ResignEvent_OrgPosPerAct`	Extracts patterns that follow the general form "Organization, Position, Person, Action"	**Performance Technologies, Inc.** (Nasdaq: PTIX), today announced that **current President and Chief Executive Officer**, **Donald L. Turrell**, will **leave** the Company's executive management at the end of 2005 to explore personal interests.

Rule	Description	Example
`ResignEvent_PerActPosOrg`	Extracts patterns that follow the general form "Person, Action, Position, Organization"	**Mr. Victor Oppleman resigned** as **president** of **MainNerve, Inc.** effective March 4, 2005.
`ResignEvent_PerActOrgBoD`	Extracts patterns that follow the general form "Person, Action, Organization, Position"	**Willy Kiekens resigned** from the **IMF**'s **board of directors**.

7.3 Extracting Product Release Events

The product release rules are designed to extract information about the announcement of new products, including the company, date, and price. The rules try to isolate novel product names and do not rely on the established set of extraction-recognized `PRODUCT` entities.

The following table describes the rules available for product release:

Rule (though labeled only as "ProductRelease")	Description	Example
`ProductRelease_OrgRelProd`	Extracts patterns that follow the general form: "Organization releases Product"	**Microsoft Corp** said it plans to **release** its **Microsoft Internet Information Service 3.0** software, a web page development package.
		TRW Automotive Holdings Corp. announced plans to **un veil** the company's **integrated safety system platform**.

Rule (though labeled only as "ProductRelease")	Description	Example
`ProductRelease_OrgAnnProd`	Extracts patterns that follow the general form: "Organization announces Product"	**Apple**(R) **today announced iTunes(R) 5**, bringing new features and a refined look to the world's most popular digital music jukebox and online music store.

7.4 Extracting Merger Information

The merger rules are designed to extract information about mergers and acquisitions.

The following table describes the rules available for merger information:

Rule	Description	Example
`BuyEvent`	Extracts patterns that follow the general form "Organization1, Buys, Organization2"	**Toys R Us** will **acquire Baby Superstore Inc** in a stock merger valued at about **$376 million**.
`SellEvent`	Extracts patterns that follow the general form "Organization1, Sells, Organization2"	The Netherlands's largest bank **ABN AMRO** said on Friday it had signed a letter of intent to **sell** its MeesPierson investment bank unit to Belgian/Dutch financial group **Fortis**.
`MergeEvent_MergePre` (reported only as MergeEvent)	Extracts patterns that follow the general form "Merge, between, Organization1, and, Organization2"	The **merger** of **AUSTEL** and **Spectrum Management Authority** was announced today.
`MergeEvent_MergePost` (reported only as MergeEvent)	Extracts patterns that follow the general form "Organization1, and, Organization2, Merge"	**Namibian Fishing Industries Ltd** (Namfish) and **Namibian Sea Products Ltd** (Namsea) are to **merge** in April.

7.5 Extracting Organizational Information

The organizational information rules are designed to extract core information about companies and organizations. The following table describes the set of rules defined for organization:

Rule	Description	Example
`OrganizationPerson`	Extracts patterns that follow the general form "Person, position at Organization" with optional Phone, Fax, Email, and URL	• **Patti J. McAtee**, Director of **CalEnergy**, **402-341-4500**. • **Motorola, Inc.**, **George Grimsrud**, **847/576-2346**.
`OrganizationPhone`	Extracts patterns that follow the general form "Organization, Phone"	• **Business Objects**, North America, **+1 800 877 2340**
`OrganizationFax`	Extracts patterns that follow the general form "Organization, Fax"	• **Marriott International** press releases are available through Company News On-Call by fax, **800-758-5804, ext. 532963**.
`OrganizationEmail`	Extracts patterns that follow the general form "Organization, Email"	• **Inxight Software Inc.**, 500 Macara Avenue, Sunnyvale, CA 94085, U.S.A., Email: **info@inxight.com**
`OrganizationURL`	Extracts patterns that follow the general form "Organization, URL"	• **Seagate**'s home page address on the World Wide Web is **http://www.seagate.com**. • These documents are all available on **Intuit**'s Web site at **http://www.intuit.com**.
`OrganizationAddress`	Extracts patterns that follow the general form "Organization, Address"	• For more information contact: **Inxight Software Inc.**, **500 Macara Avenue, Sunnyvale, CA 94085**, Tel: 555.555.1212.

Rule	Description	Example
`OrganizationLocation`	Extracts patterns that follow the general form "Organization, [located in] Place" ; "Location-[based], Organization", or "Organization, Location-[based]"	• U.S. operations of **Toyota**, Japan's largest automaker, are headquartered in **New York**. • **New York-based American Express Co.** is best known for its credit card. • **Cypress Asset Management Inc**, a **Houston-based** investment adviser
`OrganizationFounder`	Extracts patterns that follow the general form "Founder, [who founded] Organization" ; "Founder, [of] Organization" ; "Organization, [founded by] Founder" , or "Organization, Founder"	• English social reformer **Sir GEORGE WILLIAMS**, who founded the **YMCA**, was born in 1821. • **Jean-Louis Gassee**, the founder of **Be Inc**. • **Canal Plus** was founded by **Andre Rousselet**. • **Apple Computer Inc.** fired co-founder **Steven Jobs** in 1985.
`OrganizationNationality`	Extracts patterns that follow the general form "Nationality, Organization" or "Organization, [of] Nationality"	• Michelin Tyre is a unit of **France**'s **Michelin S.A**. • The Miami fast-food chain, owned by **Grand Metropolitan** of **Britain**.
`OrganizationTicker`	Extracts patterns that follow the general form "Organization, Ticker"	• **Microsoft** (**Nasdaq:MSFT**) announced this quarter's earnings.

Rule	Description	Example
`OrganizationStockPrice`	Extracts patterns that follow the general form "Organization Price"; "Organization valued at Price"; "Price is Organization stock price", or "Organization shares rose/gained/and so to Price".	• **NASDAQ: BOBJ $41.34**. • **NASDAQ: BOBJ** valued at **$41.34**. • **$41.37** is **Business Objects** stock price. • **Caterpillar**'s shares rose 0.9 percent to **$85.93**. • **Microsoft** gained 2.2 percent to **$30.33**. • **Kodak** shares were up 4.5 percent to **$45.87**.

Public Sector Content

The public sector content includes a set of rules that that you can use to extract public-sector-specific information when processing and analyzing text. It is included in and supports these languages modules:

- Arabic
- English
- Simplified Chinese

Note:
Linguistic and extraction information for the right to left languages such as, Arabic, Farsi, and Hebrew is described in a separate addendum.

Depending on the language module you are using, the public sector content provides these extraction abilities:

- Arabic, English and Simplified Chinese language modules
  - Default entity types, as documented in the *Language Modules Reference* section of this guide
  - Public-sector-specific entity types, as documented in this chapter
- English language module only:
  - Event types and relation types
  - Additional rule sets for extracting public-sector-specific types of information

Related Topics

- English: Types of Information Extracted
- Public Sector Entities–Simplified Chinese
- Language Modules Reference

8.1 English: Types of Information Extracted

When used in conjunction with the standard English language module, the public sector content lets you extract public-sector-specific entities in addition to the default entity types. The public sector content also includes rules sets that allow for the extraction of additional public-sector-related information.

8.1.1 Public Sector Content Rule Sets—English

The public sector content includes the following extraction rule sets for the English language module:

Rule Set Description	Compiled (.fsm) and Source (.rul) Files
• Action Extracts information about action and travel events	`english-tf-gov-Action.fsm` `english-tf-gov-Action.rul`
• Military Units Extracts information about military units such as teams, wings, and squadrons	`english-tf-gov-MilitaryUnits.fsm` `english-tf-gov-MilitaryUnits.rul`
• Organizational Information Extracts information about organizations	`english-tf-gov-Org.fsm` `english-tf-gov-Org.rul`
• Person-Alias Extracts information about a person's possible aliases	`english-tf-gov-PersonAlias.fsm` `english-tf-gov-PersonAlias.rul`
• Person-Appearance Extracts information about a person's appearance	`english-tf-gov-PerApp.fsm` `english-tf-gov-PerApp.rul`
• Person-Attributes Extracts information about a person's non-appearance attributes	`english-tf-gov-PerAtt.fsm` `english-tf-gov-PerAtt.rul`
• Person-Relationships Extracts information about a person's relationships	`english-tf-gov-PerRel.fsm` `english-tf-gov-PerRel.rul`
• Spatial References Extracts relative spatial references, such as distances, cardinal directions, or locations	`english-tf-gov-SpatialRef.fsm` `english-tf-gov-SpatialRef.rul`

For more details about enhancing extraction rules, refer to the *SAP BusinessObjects Data Services Text Data Processing Extraction Customization Guide*.

Related Topics

- Extracting Action Events
- Extracting Travel Events
- Extracting Military Units
- Extracting Organizational Information
- Extracting a Person's Aliases
- Extracting Information About a Person's Appearance
- Extracting Information About a Person's Attributes
- Extracting Information About a Person's Relationships
- Extracting Spatial References

8.1.2 Extracting Action Events

The `Action` rules are designed to extract information related to events involving persons or organizations and their movements, creations, and transfers. Optionally, these rules also extract `Time` and `Place` attributes for `Action` events. The following table describes the `Action` rules.

Rule	Description	Example
`Action_Buy_Active`	Extracts patterns that follow the general form "Agent, [buys] Artifact"	• **Al Qaeda** purchased **diamonds** in **Belgium**. • In **April**, **Al Qaeda insurgents** stockpiled **stolen US Army goods** in a house in **southern Bagdad**.

Rule	Description	Example
`Action_Buy_Passive`	Extracts patterns that follow the general form "Artifact, [is bought by] Agent"	• The most frightening item on Mr. Bolton's list at least that is known publicly is the large amounts of **growth media** acquired by **Saddam**, and which in such substantial quantities could only be used as cultures for growing biological agents. • **This cottage** by the lake was inherited by the **Smiths**.
`Action_Capture_Active`	Extracts patterns that follow the general form "Agent, [captures] Patient"	• THE **NARCOTICS POLICE** ARRESTED **ALFONSO CASTILLO ARMENTA** IN **MIAMI** LAST **DECEMBER**. • In **April**, **Pakistan President Gen. Pervez Musharraf** apprehended **a high ranking Taliban member**.
`Action_Capture_Passive`	Extracts patterns that follow the general form "Patient, [was captured by] Agent"	• **Fathi Subuh** was arrested by the Palestinian Authority's **Preventative Security Service** (PSS) on **July 2, 1997**.
`Action_Command_Active`	Extracts patterns that follow the general form "Agent, [commands] Organization"	• **G. W. Smith** commanded the **Army** of **Northern Virginia**. • **Pakistan President Gen. Pervez Musharraf** coordinated the **Pakistani army** efforts last **April**. • **City leaders** coordinated the **City Council**.

Rule	Description	Example
`Action_Command_Passive`	Extracts patterns that follow the general form "Organization, [is commanded by] Agent"	• The **Third Regiment, Texas Volunteer Infantry** (United States Volunteers), was commanded by **Colonel R. P. Smyth**. • **Al Qaeda** is headed by **Osama Bin Laden** since **1996**. **Al Qaeda** was created by **Osama Bin Laden**.
`Action_Command_Apposi tive`	Extracts patterns that follow the general form "Agent, [is the commander of] Organization"	• In Kuwait every company commander has a specialist advising him," said **Thomas Spoehr**, the commander of the **3rd Chemical Brigade**.
`Action_Communi cate_Meet_Active`	Extracts patterns that follow the general form "Agent, [meets] Patient"	• **President Clinton** met **Jacques Chirac** in **July 1995**.
`Action_Communi cate_Meet_Passive`	Extracts patterns that follow the general form "Patient, [is met by] Agent"	• As soon as he stepped out the plane, **Paul** was met by **FBI** agents .
`Action_Communi cate_PhoneWrite_Active`	Extracts patterns that follow the general form "Agent, [phones/writes] Patient"	• **President Clinton** called **Jacques Chirac** in **July 1995**.
`Action_Communicate_Oth er_Active`	Extracts patterns that follow the general form "Agent, [communi-cates with/contacts] Patient"	• In **December 2000**, **Osama Bin Laden** communicated with **Al-Qaeda members**. **City leaders** notified the **Iraqi Army**.
`Action_Communicate_Oth er_Passive`	Extracts patterns that follow the general form "Agent, [is contact-ed by] Agent"	• **Paul** was interviewed by **John** in **December 2007**.

Rule	Description	Example
`Action_Destroy_Active`	Extracts patterns that follow the general form "Agent, [destroys] Artifact or Patient"	• **THE NARCOTICS POLICE** DESTROYED 1,200 KG OF **COCAINE**. • In **April**, **Iraqi insurgents** attacked **5 US Army convoys**. • **Al Qaeda insurgents** ransacked **houses of private Iraqis** in search for weapons.
`Action_Destroy_Passive`	Extracts patterns that follow the general form "Artifact or Patient, [was destroyed by] Agent"	• **Atta files** destroyed by the **Pentagon**. • **US Army soldiers** were attacked by **Iraqi insurgents** in **December** in **Bagdad**. • **Pilgrims** were assaulted by **Iraqi insurgents** when they headed to the **Great Mosque** last **Thursday**. • **Saddam's Presidential Palace** was taken over by **U.S. soldiers** in **2003**.
`Action_Drive_Active`	Extracts patterns that follow the general form "Agent, [drives] Vehicle"	• **MUSTAFA** DRIVES A **BLACK SUBARU**. • **Jane** drives a **green Subaru** to work.
`Action_Drive_Passive`	Extracts patterns that follow the general form "Vehicle, [was driven by] Agent"	• That dialogue took place in a **white Peugeot** driven by **John Smith**. • **A stolen black SUV** was used by the **terrorists** in the market bomb attack.
`Action_Execute_Active`	Extracts patterns that follow the general form "Agent, [executes] Patient"	• **The Iraqi terrorist group** executed **their American prisoners**.

Rule	Description	Example
`Action_Execute_Passive`	Extracts patterns that follow the general form "Agent, [is executed by] Patient"	• **Daniel Pearl** was executed by his **captors** in **Karachi**.
`Action_Finance_Active`	Extracts patterns that follow the general form "Agent, [finances] Patient"	• In **April**, **Al Qaeda** financed **Iraqi insurgents**. • **Al Qaeda** subsidizes **various international terrorist groups**. • **Some Islamic organizations** support **terrorist groups** in **Iraq**.
`Action_Finance_Passive`	Extracts patterns that follow the general form "Agent, [is financed by] Patient"	• **Some Islamic organizations** might be financed by wealthy **Saudis**.
`Action_Hire_Active`	Extracts patterns that follow the general form "Agent, [hires] Patient"	• The **Cincinnati Reds** have hired **Bob Boo** in Major League Baseball. • In **April**, **Al Qaeda** recruited **Iraqi insurgents** to attack a Mosque. • **The local soccer federation** hired **John Brown** as acting director.
`Action_Hire_Passive`	Extracts patterns that follow the general form "Patient, [is hired by] Agent"	• **HAMAL** WAS RECRUITED BY **SHEIKH MAHMOUD MOHAMMED ALI MALIK**. • **Iraqis, Jordanians, and Iranians** were enlisted by **Al Qaeda**. • **Mohammad Ata** was recruited by **Al Qaeda** in **2000** to lead the 9/11 plane attack on **New York** and **Washington**.

Rule	Description	Example
`Action_Indict_Active`	Extracts patterns that follow the general form "Agent, [indicts] Patient"	• The **UN** indicted **Sloban Milosevic** in **2000**.
`Action_Indict_Passive`	Extracts patterns that follow the general form "Patient, [is indicted by] Agent"	• In **2000**, **Milosevic** was indicted by the **United Nations**.
`Action_Injure_Active`	Extracts patterns that follow the general form "Agent, [injures] Patient"	• **Iraqi insurgents** wounded **ten civilians** at an outdoor market in **Baghdad** last **April**.
`Action_Injure_Passive`	Extracts patterns that follow the general form "Agent, [is injured by] Patient"	• **Three men** were injured by the **Taliban** attack in **Kabul**.
`Action_Kill_Active`	Extracts patterns that follow the general form "Agent, [kills] Patient"	• **Lee Harvey Oswald** killed **President Kennedy** in **November 1963**.
`Action_Kill_Passive`	Extracts patterns that follow the general form "Patient, [is killed by] Agent"	• **Many people** were slaughtered by **Jeffrey Dahmers** in the **1980's** and **1990's**.
`Action_Make_Active`	Extracts patterns that follow the general form "Agent, [makes] Artifact"	• **Hamal** made **explosives** for money. **John** builds **IEDs**.
`Action_Make_Passive`	Extracts patterns that follow the general form "Artifact, [made by] Agent"	• **The bombs** were made by **Spanish Al Qaeda** cells.
`Action_Make_MakerOf`	Extracts patterns that follow the general form "Agent, [is an] Artifact [maker]"	• **KAMIL** IS A MORE ADVANCED **BOMB** MAKER THAN HAMAL.

Rule	Description	Example
`Action_Participate_Active`	Extracts patterns that follow the general form "Agent, [participates in/joins] Organization"	• In **April**, **Al Qaeda** cooperated with the **Taliban**. • **Mr. Roh** will join the **Uri Party** in **January 2004**.
`Action_Participate_Passive`	Extracts patterns that follow the general form "Organization, [was joined by] Agent"	• **Al Qaeda** was joined by **other Muslim extremists** in **England**. • The **United Nations** is served by **145 national delegates**.
`Action_Receive_Pay_Active`	Extracts patterns that follow the general form "Recipient, [receives] Payment"	• **Jamal Ahmed Al-Fadl** received **$10,000** for his time and effort and did not take a further role in the uranium acquisition. • **John** received **$2000** for his old car.
`Action_Receive_Pay_Passive`	Extracts patterns that follow the general form "Payment, [is received by] Recipient"	• Articles by Tina Griego showed that the largest contribution was the **$46,000** received by **Manny Aragon**. • A total of **9.5 million dollars** were incorrectly charged by **Halliburton company** to the US Army.
`Action_Survey_Active`	Extracts patterns that follow the general form "Agent, [surveys] Patient or Artifact"	• **Al Qaeda** assessed the **US Army** positions. • **Osama Bin Laden** was reviewing **high ranking Al Qaeda members**. • **The party leadership** evaluated **would-be candidates**.
`Action_Survey_Passive`	Extracts patterns that follow the general form "Patient or Artifact [is surveyed by] Agent"	• **North Korean nuclear facilities** will be inspected by the **UN Nuclear Agency**.

Rule	Description	Example
`Action_Train_Active`	Extracts patterns that follow the general form "Agent, [trains] Patient"	• Remember when the **CIA** was funding and training **Bin Laden** and his boys to fight against the Russians? • The **Montreal Baseball School** trained **John** before he played in **New York**.
`Action_Train_Passive`	Extracts patterns that follow the general form "Patient, [was trained by] Agent"	• **Spivey** trained by **Whitaker!** • **Mohammad Ata** was trained by **Al Qaeda operatives** in **Pakistan**.
`Action_Train_With`	Extracts patterns that follow the general form "Patient, [trains with] Agent"	• **John** trained with the **FBI.** • **Brigitte** trained with **Algerian al-Qaeda** in rural **France** and worked for Christian Ganczarski and Karim Mehdi, the former of whom was a lieutenant for Khalid Sheikh Mohammed (the latter was planning a Bali-style attack on Reunion Island).
`Action_Transport_Active`	Extract patterns that follow the general form "Agent, [transported] Artifact or Patient"	• **Abu Mohjen** was later indicted for his role in transporting **the arms** by sea to Kanj's bases. • In **April**, the **US Army** deployed **25,000 more soldiers** all over **Iraq**. • **John** transported **the stolen goods** across the border.

Rule	Description	Example
`Action_Transport_Passive`	Extracts patterns that follow the general form "Artifact or Patient, [was transported by] Agent"	• **Mohammed** was transported by **Al Qaeda** to **Jordan**. • Over the past **3 years**, **presumed terrorists** were transported by the **US Government** to an undisclosed location. • **Poor Mexican illegal immigrants** were deported by **US immigration officials**.

8.1.3 Extracting Travel Events

The `Action` events also include `Travel` rules that are designed to extract information about individuals and their travel events. The following table describes the `Travel` rules defined.

Rule	Description	Example
`Travel_visited`	Extracts patterns that follow the general form "Person, visited, Target Destinations." The date can appear either at the beginning or the end of the clause.	• In **April**, **Pakistan President Gen. Pervez Musharraf visited Kabul**. • **Pakistan President Gen. Pervez Musharraf visited Kabul** in **April**.

Rule	Description	Example
`Travel_visited_When`	Extracts patterns that follow the general form "Person, Date, when, visited, Target Destinations." **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **John Doe** was arrested in **Apri**, when he **visited Kabul**.
`Travel_enteredFrom`	Extracts patterns that follow the general form "Person, entered, TargetDestinations, from, Source Destination." The date can appear at the beginning, middle, or the end of the clause.	• In **December 2000**, **Mr. Benatta entered the United States** from **Canada**. • **Mr. Benatta entered the United States** from **Canada** in **December 2000**.
`Travel_enteredFrom_When`	Extracts patterns that follow the general form "Person, Date, when, entered, TargetDestinations, from, Source Destination." **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **John Doe** was arrested in **December 2000**, when he **entered the United States** from **Canada**.
`Travel_arrivedInFrom`	Extracts patterns that follow the general form "Person, arrived, in, TargetDestinations, from, Source Destination." The date can appear at the beginning, middle, or the end of the clause.	• On **Christmas Day**, **Queen Elizabeth arrived** in **San Francisco** from **England**.

Rule	Description	Example
`Travel_arrivedIn_When`	Extracts patterns that follow the general form "Person, Date, when, arrived, in, TargetDestinations, from, Source Destination" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **John Doe** was arrested on **Christmas Day**, when he **arrived** in **San Francisco** from **England**.
`Travel_cameToFrom`	Extracts patterns that follow the general form "Person, came back to, Target Destinations, from, Source Destination." The date can appear at the beginning, middle, or the end of the clause.	• On **Dec. 1**, **Clinton flew** back to **the United States** from **Germany**.
`Travel_cameToFrom_When`	Extracts patterns that follow the general form "Person, Date, when, came, back to, Target Destinations, from, Source Destination" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **John Doe** was arrested on **Dec. 1**, when he **flew** back to **the United States** from **Germany**.

Rule	Description	Example
`Travel_cameFromTo`	Extracts patterns that follow the general form "Person, came from, Source Destination, back to, Target Destinations." The date can appear at the beginning, middle, or the end of the clause.	• On **Dec. 1**, **Clinton flew** from **Germany** back to **the United States**.
`Travel_cameFromTo_When`	Extracts patterns that follow the general form "Person, Date, when, came, from, Source Destination, back to, Target Destinations" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **John Doe** was arrested on **Dec. 1**, when he **flew** from **Germany** back to **the United States**.
`Travel_departedFor`	Extracts patterns that follow the general form "Person, departed, Source Destination, for, Target Destination." The date can appear at the beginning, middle, or the end of the clause.	• On **Christmas**, **Mohamed departed Germany** for **the Netherlands**.
`Travel_departedFor_When`	Extracts patterns that follow the general form "Person, Date, when, departed, Source Destination, for, Target Destinations" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **John Doe** was arrested on **Christmas Eve**, when he **departed Germany** for **the Netherlands**.

Rule	Description	Example
`Travel_gainedEntry IntoFrom`	Extracts patterns that follow the general form "Person, gained entry, to, Target Destinations, from, Source Destination." The date can appear at the beginning, middle, or the end of the clause.	• In **1992**, **Mohamad Hammoud gained entry** to **the U.S.** through **Venezuela**.
`Travel_gainedEntry IntoFrom_When`	Extracts patterns that follow the general form "Person, Date, when, gained entry, to, Target Destinations, from, Source Destination" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **John Doe** was arrested on **Oct. 10**, when he **gained entry** into **Thailand** from **Laos**.
`Travel_crossedIntoFrom`	Extracts patterns that follow the general form "Person, crossed, the border, to, Target Destinations, from, Source Destination." The date can appear at the beginning, middle, or the end of the clause.	• On **Oct. 10**, **Hambali crossed** into **Thailand** from **Laos**.

Rule	Description	Example
`Travel_crossed IntoFrom_When`	Extracts patterns that follow the general form "Person, Date, when, crossed, the border, to, Target Destinations, from, Source Destination" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **John Doe** was arrested on **Oct. 10**, when he **crossed** into **Thailand** from **Laos**.
`Travel_crossedOverToFrom`	Extracts patterns that follow the general form "Person, crossed over, to, Target Destinations, from, Source Destination." The date can appear at the beginning, middle, or the end of the clause.	• In **April**, **Mr. Singh crossed over** into **Pakistan** from **Afghanistan**.
`Travel_crosse dOverToFrom_When`	Extracts patterns that follow the general form "Person, Date, when, crossed over, to, Target Destinations, from, Source Destination" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **Mr. Singh** was arrested in **April**, when he **crossed over** into **Pakistan** from **Afghanistan**.

Rule	Description	Example
`Travel_crossedOverFromTo`	Extracts patterns that follow the general form "Person, crossed over, from, Source Destination, to, Target Destinations." The date can appear at the beginning, middle, or the end of the clause.	• In **April**, **Mr. Singh crossed over** from **Afghanistan** into **Pakistan**.
`Travel_crossedOverFromTo_When`	Extracts patterns that follow the general form "Person, Date, when, crossed over, from, Source Destination, to, Target Destinations" **Note:** For this rule to work correctly, Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **Mr. Singh** was arrested in **April**, when he **crossed over** from **Afghanistan** into **Pakistan**.
`Travel_escapedToFrom`	Extracts patterns that follow the general form "Person, escaped, to, Target Destinations, from, Source Destination." The date can appear at the beginning, middle, or the end of the clause.	• In **1980**, **Adnan Al-Ghoul escaped** to **Lebanon** from **Israel**.
`Travel_escapedToFrom_When`	Extracts patterns that follow the general form "Person, Date, when, escaped, to, Target Destinations, from, Source Destination" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **Adnan Al-Ghoul** had been on the wanted list since **1980**, when he **escaped** to **Lebanon** from **Israel**.

Rule	Description	Example
`Travel_escapedFromTo`	Extracts patterns that follow the general form "Person, escaped, from, Source Destination, to, Target Destinations." The date can appear at the beginning, middle, or the end of the clause.	• In **1980**, **Adnan Al-Ghoul escaped** from **Israel** to **Lebanon**.
`Travel_escapedFrom To_When`	Extracts patterns that follow the general form "Person, Date, when, escaped, from, Source Destination, to, Target Destinations" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **Adnan Al-Ghoul** had been on the wanted list since **1980**, when he **escaped** from **Israel** to **Lebanon**.
`Travel_fledFor`	Extracts patterns that follow the general form "Person, fled, Source Destination, for, Target Destinations." The date can appear at the beginning, middle, or the end of the clause.	• On **August 6, 1998**, **Abdullah Ahmed Abdullah fled Nairobi** for **Karachi**.
`Travel_fledFor_When`	Extracts patterns that follow the general form "Person, Date, when, fled, Source Destination, for, Target Destinations" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **Broderick** was arrested on **March 14, 2002** when he **fled** for **Canada** at the International Bridge.

Rule	Description	Example
`Travel_triedToCross IntoFrom`	Extracts patterns that follow the general form "Person, tried to cross, the border, to, Target Destinations, from, Source Destination." The date can appear at the beginning, middle, or the end of the clause.	• In **1999**, **Captain Aleksey Konkov tried to cross** into **Laredo** from **Mexico**.
`Travel_triedToCross IntoFrom_When`	Extracts patterns that follow the general form "Person, Date, when, tried to cross, the border, to, Target Destinations, from, Source Destination" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **Broderick** was arrested on **March 14, 2002** when he **tried to cross** into **Canada** at the International Bridge.
`Travel_attemptedEntry IntoFrom`	Extracts patterns that follow the general form "Person, attempted entry, to, Target Destinations, from, Source Destination." The date can appear at the beginning, middle, or the end of the clause.	• Last **October**, **Mr. Hammoud sought entry** into **Canada** from **the United States**.

Rule	Description	Example
`Travel_attemptedEntryIntoFrom_When`	Extracts patterns that follow the general form "Person, Date, when, attempted entry, to, Target Destinations, from, Source Destination" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **Broderick** was arrested on **March 14, 2002** when he **sought entry** into **Canada** at the International Bridge.
`Travel_attemptedToEnter`	Extracts patterns that follow the general form "Person, attempted to enter, Target Destinations." The date can appear either at the beginning or the end of the clause.	• In **1998**, **Abdulla Ocalan attempted to enter Italy**.
`Travel_attemptedToEnter_When`	Extracts patterns that follow the general form "Person, Date, when, attempted to enter, Target Destinations" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the travel event is applied to the antecedent of he and the date information is lost.	• **Broderick** was arrested on **March 14, 2002** when he **tried to enter Canada** at the International Bridge.

8.1.4 Extracting Military Units

The `MilitaryUnit` rule is designed to extract expressions that refer to military units, as described in the following table.

Rule	Description	Example
`MilitaryUnit`	This rule extracts long and often coordinated expressions that refer to military units.	• Paul serves in **Company A, 1st Battalion, 22nd Infantry Regiment, 1st Brigade Combat Team, 4th Infantry Division**. • Soldiers from **Company C, 3rd Battalion, 67th Armor Regiment, 4th Brigade Combat Team, 101st Airborne Division**, detained four kidnappers in Baghdad July 31.

8.1.5 Extracting Organizational Information

The organizational information rules are designed to extract core information about companies and organizations. The following table describes the set of rules defined for organization:

Rule	Description	Example
`OrganizationPerson`	Extracts patterns that follow the general form "Person, position at Organization" with optional Phone, Fax, Email, and URL	• **Patti J. McAtee**, Director of **CalEnergy**, **402-341-4500**. • **Motorola, Inc.**, **George Grimsrud**, **847/576-2346**.
`OrganizationPhone`	Extracts patterns that follow the general form "Organization, Phone"	• **Business Objects**, North America, **+1 800 877 2340**

Rule	Description	Example
`OrganizationFax`	Extracts patterns that follow the general form "Organization, Fax"	• **Marriott International** press releases are available through Company News On-Call by fax, **800-758-5804, ext. 532963**.
`OrganizationEmail`	Extracts patterns that follow the general form "Organization, Email"	• **Inxight Software Inc.**, 500 Macara Avenue, Sunnyvale, CA 94085, U.S.A., Email: **info@inxight.com**
`OrganizationURL`	Extracts patterns that follow the general form "Organization, URL"	• **Seagate**'s home page address on the World Wide Web is **http://www.seagate.com**. • These documents are all available on **Intuit**'s Web site at **http://www.intuit.com**.
`OrganizationAddress`	Extracts patterns that follow the general form "Organization, Address"	• For more information contact: **Inxight Software Inc.**, **500 Macara Avenue, Sunnyvale, CA 94085**, Tel: 555.555.1212.
`OrganizationLocation`	Extracts patterns that follow the general form "Organization, [located in] Place" ; "Location-[based], Organization", or "Organization, Location-[based]"	• U.S. operations of **Toyota**, Japan's largest automaker, are headquartered in **New York**. • **New York-based American Express Co.** is best known for its credit card. • **Cypress Asset Management Inc**, a **Houston-based** investment adviser

Rule	Description	Example
`OrganizationFounder`	Extracts patterns that follow the general form "Founder, [who founded] Organization" ; "Founder, [of] Organization" ; "Organization, [founded by] Founder" , or "Organization, Founder"	• English social reformer **Sir GEORGE WILLIAMS**, who founded the **YMCA**, was born in 1821. • **Jean-Louis Gassee**, the founder of **Be Inc**. • **Canal Plus** was founded by **Andre Rousselet**. • **Apple Computer Inc.** fired co-founder **Steven Jobs** in 1985.
`OrganizationNationality`	Extracts patterns that follow the general form "Nationality, Organization" or "Organization, [of] Nationality"	• Michelin Tyre is a unit of **France**'s **Michelin S.A**. • The Miami fast-food chain, owned by **Grand Metropolitan** of **Britain**.
`OrganizationTicker`	Extracts patterns that follow the general form "Organization, Ticker"	• **Microsoft** (**Nasdaq:MSFT**) announced this quarter's earnings.
`OrganizationStockPrice`	Extracts patterns that follow the general form "Organization Price"; "Organization valued at Price"; "Price is Organization stock price", or "Organization shares rose/gained/and so to Price".	• **NASDAQ: BOBJ $41.34**. • **NASDAQ: BOBJ** valued at **$41.34**. • **$41.37** is **Business Objects** stock price. • **Caterpillar**'s shares rose 0.9 percent to **$85.93**. • **Microsoft** gained 2.2 percent to **$30.33**. • **Kodak** shares were up 4.5 percent to **$45.87**.

8.1.6 Extracting a Person's Aliases

The `PersonAlias` rules are designed to extract information about individuals and the possible alternate names and aliases they might use. The following table describes the rules defined for `PersonAlias`.

Rule	Description	Example
`PersonAlias_alias_Person`	Extracts patterns that follow the general form "Person, alias, Person or Proper Noun, Punctuation"	• **Qasim Bokhari**, a/k/a **Syed Qasim Ali Bokhari**, and a/k/a **Kasim Bokhari.**
`PersonAlias_alias_Proper`	Extracts patterns that follow the general form "Proper Noun, alias, Person, Punctuation"	• **Sam Sneed**, a.k.a. **William Smith**, is one of the group.
`PersonAlias_alias_PersonParen`	Extracts patterns that follow the general form "Open Parenthesis, alias, Person or Proper Noun, Close Parenthesis"	• **Qasim Bokhari** (a/k/a **Syed Qasim Ali Bokhari**, and a/k/a **Kasim Bokhari**).
`PersonAlias_alias_ProperParen`	Extracts patterns that follow the general form "Proper Noun, Open Parenthesis, alias, Person, Close Parenthesis"	• **Sam Sneed**, (a.k.a. **William Smith**), is one of the group.
`PersonAlias_or`	Extracts patterns that follow the general form "Person, Comma, or, Person, Comma"	• Soldiers spotted the groups led by another Abu Sayyaf leader, **Mr. Umbra Jumdail**, or **Dr. Abu Pula**, as he's known.
`PersonAlias_AlsoKnownAs`	Extracts patterns that follow the general form "Person, also known as, Person or Proper Noun"	• **Nasr Fahmi Nasr Hassanein**, known as **Mohamed Salah**, was also involved.

Rule	Description	Example
`PersonAlias_AlsoKnownAs_Who`	Extracts patterns that follow the general form "Person, who is, also know as, Person or Proper Noun" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the alias found is applied to who and its antecedent.	• He left Egypt in 1989 and went to Peshawar, where he met **Dr. Rashid Abd-al-Al-im**, who is known as **Dr. Fadl**.
`PersonAlias_AlsoKnownAs_NPWho`	Extracts patterns that follow the general form "Person, Comma, NP, who is, also know as, Person or Proper Noun"	• **ABU SUHAYB AL-AMRIKI**: A 25-year-old U.S. citizen who is known as **Adam Pearlman**.
`PersonAlias_AlsoKnownAs_Quote`	Extracts patterns that follow the general form "Person, also known as, Open Quote, General Alias, Close Quote"	• **Ahmed Khalhan Ghailiani**, also known as "**Ahmed the Tanzanian**," "**Foopie**," and "**Fupi**."
`PersonAlias_UsingTheName`	Extracts patterns that follow the general form "Person, using the name, Person or Proper Noun"	• **Filippo Bertotti**, using the nickname **Filippo Rumi**, has written several articles in the newspaper Il Manifesto.
`PersonAlias_UsingTheName_Who`	Extracts patterns that follow the general form "Person, who is, using the name, Person or Proper Noun" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the alias found is applied to who and its antecedent.	• **Subhi Abdel-Aziz El-Gohari Abu Sittah**, who also goes by the name **Abu Hafs El-Masri**.

Rule	Description	Example
`PersonAlias_UsingTheName_NPWho`	Extracts patterns that follow the general form "Person, Comma, NP, who is, using the name, Person or Proper Noun"	• **ABU SUHAYB AL-AMRIKI**: A 25-year-old U.S. citizen who is using the name of **Adam Pearlman**.
`PersonAlias_UsingTheName_Quote`	Extracts patterns that follow the general form "Person, using the name, Open Quote, General Alias, CloseQuote"	• **Filippo Bertotti**, using the nickname "**Filippo Rumi**", has written several articles in the newspaper Il Manifesto.
`PersonAlias_UsingTheNames`	Extracts patterns that follow the general form "Person, using the names, Person or Proper Noun"	• **MICHAEL KAIGHN** is using the aliases **D.S.** and **Patrick Grogan**.
`PersonAlias_UsingTheNames_Who`	Extracts patterns that follow the general form "Person, who is, using the names, Person or Proper Noun" **Note:** For this rule to work correctly, advanced parsing must be turned off. Otherwise, if advanced parsing is on, the alias found is applied to who and its antecedent.	• **Subhi Abdel-Aziz El-Gohari Abu Sittah**, who also goes by the names **Abu Hafs El-Masri** and **Mohamed Atif**.
`PersonAlias_UsingTheNames_NPWho`	Extracts patterns that follow the general form "Person, Comma, NP who is, using the names, Person or Proper Noun"	• **ABU SUHAYB AL-AMRIKI**: A 25-year-old U.S. citizen who also goes by the names **Adam Pearlman** and **Adam Gadahn**.
`PersonAlias_UsingTheNames_Quote`	Extracts patterns that follow the general form "Person, using the names, Open Quote, General Alias, CloseQuote"	• **MICHAEL KAIGHN** is using the aliases "**D.S.**" and "**Patrick Grogan**".

Rule	Description	Example
`PersonAlias_AlsoKnownAsPRE`	Extracts patterns that follow the general form "also known as, Person or Proper Noun, Comma, Person"	• Also known as **Robert**, **John** is the leader of the organization.
`PersonAlias_AlsoKnownAsPRE_Quote`	Extracts patterns that follow the general form "also known as, Open Quote, General Alias, Close Quote, Comma, Person"	• Using the alias "**Jafar the Pilot**", **John** is the leader of the organization.

8.1.7 Extracting Information About a Person's Appearance

This person-appearance (`PerApp`) rules are designed to extract appearance-related attributes of a person, such as height and eyes color, or style of dress. The following table describes the rules defined for `PerApp`.

Rule	Description	Example
`PerApp_Age`	Extract patterns that follow the general form "Person, Age [years old]"	• **Mary** is a **32-year-old** registered nurse.
	Extract patterns that follow the general form "Person, Age"	• **Febe Velazquez**, **age 27** and the mother of three, died from severe skull injuries.
`PerApp_Dress`	Extract patterns that follow the general form "Person [wears], clothing"	• **Hamal** wears **t-shirt and jeans**.

Rule	Description	Example
`PerApp_Eyes`	Extract patterns that follow the general form "Person, [has] Eyes"	• **Ahmad** has piercing **brown** eyes.
	Extract patterns that follow the general form "Person, Eyes [are]"	• **Jane**'s eyes are **green.**
`PerApp_Hair`	Extract patterns that follow the general form "Person, [has] Hair"	• **JOHN** HAS **MEDIUM LENGTH BLACK** HAIR.
	Extract patterns that follow the general form "Person, hair [is] Hair color or shape"	• **Jane**'s hair is **brown**.
`PerApp_Height`	Extract patterns that follow the general form "Person, Height"	• **Muhammed** is 35 years old, approximately **186 cm** tall, weighs 70 kg, has a thin build.
`PerApp_Weight`	Extract patterns that follow the general form "Person, Weight"	• **AMIR ((AL-JALIL**)), IS 43 YEARS OLD, 170 CENTIMETERS TALL, **90 KILOGRAMS** IN WEIGHT, WITH AN OVERWEIGHT BUILD.
	Extract patterns that follow the general form "Person, weighs [Weight]"	• **John** weighs approximately **190 pounds**.

8.1.8 Extracting Information About a Person's Attributes

This person-attributes (`PerAtt`) rules are designed to extract non-appearance-related attributes of a person. The following table describes the rules defined for `PerAtt`.

Rule	Description	Example
PerAtt_Address	Extract patterns that follow the general form "Person, [lives at] address"	• **JOSE IGNACIO RODRIGUEZ** RESIDES AT **1500 LARREA ST**.
	Extract patterns that follow the general form "Person, [lives in place] address"	• When **John** left his flat at **112 Boulevard John Kennedy** in Corbeil just outside Paris in October 1997, he was heading for London.
	Extract patterns that follow the general form "Person, [at] address"	• You can write to **Bill Smith** at **7701 Boca Ciega Drive, St. Pete Beach, Florida 33706**.
PerAtt_Location	Extract patterns that follow the general form "Person, place" inside a clause	• **John** lived in **Montreal.**
	Extract patterns that follow the general form "Person, place"	• **Bah**'s rental house in **Mon rovia.**
	Extract patterns that follow the general form "Person, place [resident]"	• **JOHN SMITH**, A **MIAMI** RESIDENT.
	Extract patterns that follow the general form "Person, [resident of] place"	• **LUZ MERY GARCIA**, A RESIDENT OF **SANTO DOMINGO**.
PerAtt_Location_verbSn	Extract patterns that follow the general form "Person, place" inside a sentence. **Note:** For this rule to work correctly, advanced parsing must be turned off.	• **Saad Khayyat**, who has lived in **New Zealand** for more than seven years, said the war was about the Middle East being "remade" to suit American interests.

Rule	Description	Example
`PerAtt_Phone`	Extract patterns that follow the general form "Person['s number is], Phone"	• **John**'s phone number is **888-333-1212**.
	Extract patterns that follow the general form "Person, Phone"	• **Roderick Liddell** (**telephone: (0)3 88 41 24 92**)
`PerAtt_Phone_verbSn`	Extract patterns that follow the general form "[contact] Person, [phone] Phone"	• You can contact **Karl Horwitz** in Paris for details and prices (**Phone: 47 42 17 11**; FAX: 47-42-80-44).
`PerAtt_Vehicle`	Extract patterns that follow the general form "Person, [drives] Vehicle"	• **John** drives a **red Subaru**.

Rule	Description	Example
PerAtt_Nationality	Extract patterns that follow the general form "Person, [is] Nationality"	• **John** is **Canadian.**
	Extract patterns that follow the general form "Person, [is from] Nationality"	• **John** is from **Canada.**
	Extract patterns that follow the general form "Person['s nationality is], Nationality"	• **John** 's nationality is **Canadian.**
	Extract patterns that follow the general form "Nationality, [citizen], Person"	• MEXICAN NARCOTICS POLICE HAVE ARRESTED **COLOMBIAN** CITIZEN **JORGE HUMBERTO CHALARIA**.
	Extract patterns that follow the general form "Person, Nationality [citizen]"	• **JENNIFER CASOLO**, A **U.S.** CITIZEN WAS ARRESTED BY THE SECURITY CORPS.
	Extract patterns that follow the general form "Person, [citizen of] Nationality"	• **Marwan al-Shehhi**, a citizen of the **United Arab Emirates**.

Rule	Description	Example
PerAtt_Affiliation	Extract patterns that follow the general form "Person, [join] Organization"	• **John** joined the **Democratic Party**.
	Extract patterns that follow the general form "Person, [a member of] Organization"	• **Rep. James Walsh**, a senior member of the **Appropriations Committee**.
	Extract patterns that follow the general form "Organization [member], Person"	• **M-19 DIRECTORATE** MEMBER **RAFAEL VERGARA**.
	Extract patterns that follow the general form "Person, an Organization [member]"	• **Yael Dayan**, a **Labour Party** member
PerAtt_Occupation	Extract patterns that follow the general form "Person, [is] Occupation"	• **John** is a **carpenter.**
	Extract patterns that follow the general form "Person, [works] Occupation"	• **Robert Fisk** has worked as a **journalist** in the Middle East for years.
	Extract patterns that follow the general form "Person, [a] Occupation"	• No water, no electricity, no work, no medicine, said **Ali Noor**, a **retired chemical engineer**.
PerAtt_Possession	Extract patterns that follow the general form "Person, [has] Artifact"	• **John** has a **bike.**
	Extract patterns that follow the general form "Artifact, [owned by] Person"	• The **bike** owned by **John.**

8.1.9 Extracting Information About a Person's Relationships

The person-relationships (`PerRel`) rules are designed to extract relationships between two people, whether familial, social, or work-related. The following table describes the rules defined for `PerRel`.

Rule	Description	Example
`PerRel_ParentChild`	Extract patterns that follow the general form "Parent, Child"	• **Surjit Kaur**'s son **Prabjot Singh**. • **Surjit Kaur**'s son, **Prabjot Singh**. • **Fateha Gazi**, the mother of **Rana and Amy**. • **Surjit Kaur** is **Prabjot Singh**'s father.
	Extract patterns that follow the general form "Child, Parent"	• **Prabjot Singh**'s father **Surjit Kaur**. • **Prabjot Singh**'s father, **Surjit Kaur**. • **Rana**, the daughter of **Fateha Gazi**. • **Prabjot Singh** is **Surjit Kaur**'s son. • **Paul and John Singh** are **Surjit Kaur**'s sons.
`PerRel_Sibling`	Extract patterns that follow the general form "Sibling, Sibling"	• **Mohamed Kubwa**'s half-brother **Amina**. • **Mohamed Kubwa**'s half-brother, **Amina**. • **Amina**, half-brother of **Mohamed Kubwa**. • **Mohamed Kubwa** and his half-brother **Amina**. • **John** enlisted his brother **Joe**.

Rule	Description	Example
`PerRel_Spouse`	Extract patterns that follow the general form "Spouse, Spouse"	• **John Manningham**'s wife **Patricia**. • **John Manningham'**s wife, **Patricia**. • **Patricia**, the wife of **John Manningham**. • **John Manningham** and his wife **Patricia**. • **John** asked his wife **Donna** to help him. • Bach's father, **Johann Ambrosius**, married **Maria Elisabeth Lämmerhirt** on April 8, 1668.
`PerRel_Relative`	Extract patterns that follow the general form "Relative, Relative"	• **Bin Laden**'s brother-in-law **Mohammed Jamal Khalifa**. • **Bin Laden**'s brother-in-law, **Mohammed Jamal Khalifa**. • **John** talked to his uncle **Charlie**.
`PerRel_Associate`	Extract patterns that follow the general form "Associate, Associate"	• Italian acting Finance Minister **Giovanni Goria** met Treasury Secretary **James Baker** last night. • **John** and **Fred** met at the park. • A curious story was told by **Swaran Singh** and his friend **Sukhdev Singh**. • **John** and **Fred** are roommates.

8.1.10 Extracting Spatial References

The `SpatialReference` rules are designed to extract spatial references that can be either exact based on distance, direction, and place (30 KM South of Kandahar), or vague based on preposition and place (in Kandahar, near Kandahar), as described in the following table.

Rule	Description	Example
`SpatialReference_Exact`	This rule extracts exact spatial references that contain distance, direction, and place, as in "n km/miles North/South/West/East of Place".	• **15 km north** of **Kandahar** • **South** of **Kabul**
`SpatialReference_Vague`	This rule extracts vague spatial references that are made of a preposition and a place, as in "in/near Place".	• **near Mosul** • **in** the **vicinity of Kandahar** • **near Mosul, Kirkuk, and Basrah**

8.2 Simplified Chinese: Types of Information Extracted

When used in conjunction with the standard Simplified Chinese language module, the public sector content lets you extract public-sector-specific entities in addition to the default entity types.

8.2.1 Public Sector Entities–Simplified Chinese

In addition to extracting all of the entity types found by the standard Simplified Chinese module, the public sector content has the following behaviors:

- All common mention entity types are extracted by default.
- In addition to the standard entity types, the public sector content includes these specialized entity types: VEHICLE, and WEAPON.
- In addition to the standard common mention types, it also includes these specialized common mention entity types: COMMON_VEHICLE, and COMMON_WEAPON.

8.2.1.1 VEHICLE

Methods of transportation, extracted as one of the following subtypes:

- `AIR`–Air vehicles, such as airplanes and helicopters:
  - 波音767
  - 空中客车
- `LAND`–Land vehicles, including the color, year, model and make of the vehicle:
  - 丰田汽车
  - 凯迪拉克
  - 切诺基吉普车
- `WATER`–Water vehicles:
  - 泰坦尼克
  - 永丰舰

8.2.1.2 WEAPON

Weapons, extracted as one of the following subtypes:

- `BIOLOGICAL`–Names of bacteria, viruses, fungi, natural toxins, and diseases that have been officially identified as used to harm humans, plants (crops), and animals, or as potential biological threats. This also extracts entities that describe the means for the dispersal of any of these weapons:
  - 炭疽武器
- `BLUNT`–Names of weapons designed or used as bludgeoning instruments:
  - X 号钝器
- `CHEMICAL`–Names of chemical substances that have been officially identified as used to harm humans, plants (crops), and animals, or as potential chemical threats. This also extracts entities that describe the means of the dispersal of any of these weapons:
  - 芥子毒气
- `EXPLODING`–Names of substances that cause damage by exploding:
  - TNT炸药

- NUCLEAR–Names of weapons that have been officially identified as used to harm humans, plants (crops), and animals through the dispersal of radiological or nuclear energies, or have been identified as potential nuclear threats:
  - 长崎原子弹
  - 广岛原子弹
- PROJECTILE–Names of weapons that are designed or used to be projected at great speed for the purpose of causing damage:
  - 响尾蛇导弹
  - 赤链蛇飞弹
- SHARP–Names of weapons designed or used to cut, slash, jab, and hack:
  - 碧血剑
- SHOOTING–Names of weapons that are designed or used to send projectile objects at great speed for the purpose of causing damage:
  - 卡宾枪
  - 来福枪
- OTHER–Names of weapons that do not fit into a specific subtype:
  - X型武器

8.2.1.3 COMMON_VEHICLE

Methods of transportation, extracted as one of the following subtypes:

- AIR–Air vehicles, such as airplanes and helicopters:
  - 战斗机
  - 运载火箭
  - 飞船
- LAND–Land vehicles, including the color, year, model and make of the vehicle:
  - 面包车
  - 坦克
  - 出租车
- WATER–Water vehicles:
  - 快艇

- 军舰
- 航空母舰

8.2.1.4 COMMON_WEAPON

Weapons, extracted as one of the following subtypes:

- `BIOLOGICAL`–Names of bacteria, viruses, fungi, natural toxins, and diseases that have been officially identified as used to harm humans, plants (crops), and animals, or as potential biological threats. This also extracts entities that describe the means for the dispersal of any of these weapons:
  - 生物武器
  - 细菌炸弹
- `BLUNT`–Names of weapons designed or used as bludgeoning instruments:
  - 钝器
  - 警棍
- `CHEMICAL`–Names of chemical substances that have been officially identified as used to harm humans, plants (crops), and animals, or as potential chemical threats. This also extracts entities that describe the means of the dispersal of any of these weapons:
  - 化学武器
  - 催泪弹
- `EXPLODING`–Names of substances that cause damage by exploding:
  - 炸弹
  - 炸药
  - 雷管
- `NUCLEAR`–Names of weapons that have been officially identified as used to harm humans, plants (crops), and animals through the dispersal of radiological or nuclear energies, or have been identified as potential nuclear threats:
  - 核导弹
  - 氢弹
  - 原子弹
- `PROJECTILE`–Names of weapons that are designed or used to be projected at great speed for the purpose of causing damage:
  - 导弹
  - 飞弹

- `SHARP`–Names of weapons designed or used to cut, slash, jab, and hack:
  - 宝剑
  - 匕首
  - 利斧
- `SHOOTING`–Names of weapons that are designed or used to send projectile objects at great speed for the purpose of causing damage:
  - 手枪
  - 冲锋枪
  - 半自动步枪
- `OTHER`–Common nouns of weapons that do not fit into a specific subtype:
  - 武器
  - 凶器
  - 军火

Common Mentions Content

Common mentions content includes a set of entity types and rules that address requirements for extracting common mentions. You can use this content to extract these specific types of information when processing and analyzing text:

Rule Set Description	Compiled (.fsm) and Source (.rul) Files
• Common Mentions Extracts information about common noun mentions	`english-tf-com.fsm` `english-tf-com.rul`

Note:
The common mentions content is included in and supports the English language module only.

For details about using common mentions content to enhance extraction rules, refer to the *SAP BusinessObjects Data Services Text Data Processing Extraction Customization Guide*.

9.1 Common Noun Mentions

Common noun mentions refers to the use of common nouns to refer to entities such as organizations, persons, or facilities which would normally also be referred to by proper nouns. They are defined as noun phrases headed by an appropriate noun. Both singular and plural forms are matched. Proper nouns and modifiers are also included.

This section describes the common noun mentions supported by the English language module and examples of each. Click each link to jump to that subsection: COMMON_ADDRESS1 and COMMON_ADDRESS2, COMMON_CONTINENT, COMMON_COUNTRY, COMMON_FACILITY, COMMON_GEO_AREA, COMMON_GEO_FEATURE, COMMON_LOCALITY, COMMON_ORGANIZATION, COMMON_PERSON, COMMON_PRECURSOR, COMMON_REGION,COMMON_VEHICLE, COMMON_WEAPON.

Entity Type and Description	English
COMMON_ADDRESS1 Common names for addresses	X
COMMON_ADDRESS2 Common names for second part of addresses	X
COMMON_CONTINENT Common names for continents	X
COMMON_COUNTRY Common names for countries including common nouns for geo-political entities for which the conventional labels do not apply, such as disputed territories or territories that have not been internationally recognized	X
COMMON_FACILITY Common names for man-made structures	X
COMMON_GEO_AREA Common names for geographical regions, districts, states, and provinces	X
COMMON_GEO_FEATURE Common names for places that are not geographical or political regions	X
COMMON_LOCALITY Common names for cities	X
COMMON_ORGANIZATION Common names for organizations	X

Entity Type and Description	English
`COMMON_PERSON` Common names for persons	X
`COMMON_PRECURSOR` Common names for weapon precursors	X
`COMMON_REGION` Common names of counties, prefectures, districts, and so on	X
`COMMON_VEHICLE` Common methods of transportation, of various subtypes	X
`COMMON_WEAPON` Common names for weapons, of various subtypes	X

9.1.1 COMMON_ADDRESS1 and COMMON_ADDRESS2

Common nouns for addresses, mirroring their named counterparts even if there is not any real difference in content between them, for example:

- fictitious address

9.1.2 COMMON_CONTINENT

Common nouns for the entirety of any continent, for example:

- major continents

9.1.3 COMMON_COUNTRY

Common nouns for the entirety of any country. This list also includes common nouns for geo-political entities for which the conventional labels do not apply, such as disputed territories or territories that have not been internationally recognized, for example:

- beloved motherland
- major countries
- Native American reservation
- smaller nations

9.1.4 COMMON_FACILITY

Common nouns for man-made structures, extracted as one of the following subtypes:

- `AIRPORTS`–The names of primarily man-made or man-maintained structures whose primary use is as air transportation terminals. For example,
  - commercial airport
  - busy air field
  - public heliport
- `BUILDGROUNDS`–The names of architectural and civil engineering structures, and outdoor spaces that are mainly man-made or man-maintained. There is no distinction with respect to their function, they could be civil or military facilities, they could be used for work or entertainment, or they could be monuments. For example,
  - public library
  - famous national archives
  - national park
  - training camp
  - train station
  - naval port
- `PATH`–The names of primarily man-made or man-maintained structures that allows fluids, energy, persons, animals, or vehicles to pass from one location to another. For example,
  - deserted street

  - narrow canal
  - heavily defended bridge
- `PLANT`–The names of facilities composed by one or more buildings used for industrial purposes. For example,
  - oil refinery
  - copper smelter
  - thermal power station
  - steel foundry
- `SUBAREA`–The names of portions of facilities, typically architectural ones, that are able to contain people, animals, or objects. For example,
  - small atrium
  - cold cellar
  - new kitchen
  - top-floor apartment

9.1.5 COMMON_GEO_AREA

Common nouns for geographical regions that are not political entities or natural locations, extracted as one of the following subtypes:

- `DOMESTIC`–Common nouns for locations that do not cross national borders, for example:
  - remote region
  - open frontier area
- `INTL`–Common nouns for locations that cross international borders, for example:
  - overseas

9.1.6 COMMON_GEO_FEATURE

Common nouns for natural geographical or political regions, extracted as one of the following subtypes:

- `BOUNDARY`–Common nouns for locations such as a border, for example:
  - northern border

    - unaccessible frontiers
- `CELESTIAL`–Common nouns for locations outside of Earth, for example:
    - largest planet
    - night sky
- `LAND`–Common nouns for geologically or ecosystemically designed non-artificial locations, for example:
    - mountain range
    - French seaside
- `WATER`–Common nouns for bodies of water, for example:
    - saltwater lake
    - flooding rivers

9.1.7 COMMON_LOCALITY

Common nouns for cities, for example:

- border town
- densely populated cities

9.1.8 COMMON_ORGANIZATION

Common nouns for organizations, extracted as one of the following subtypes:

- `COMMERCIAL`–Common nouns for companies, for example:
    - small robotics company
    - pesticides manufacturers
    - world's fourth-biggest airline
- `EDUCATIONAL`–Common nouns for institutions focused on education, for example:
    - private university
    - public colleges
- `ENTERTAINMENT`–Common nouns for institutions focused on entertainment, for example:

  - contemporary circus
  - theater company
- `GOVERNMENT`–Common nouns for institutions related to government, politics, or the state, for example:
  - Obama government
  - battalion
  - justice system
- `MEDIA`–Common nouns for institutions related to the media, for example:
  - news service
  - television station
- `MEDICALSCIENCE`–Common nouns for institutions related to medicine or research, for example:
  - health group
  - teaching hospital
- `RELIGIOUS`–Common nouns for institutions related to religion, for example:
  - Catholic church
  - powerful archdiocese
- `SPORTS`–Common nouns for institutions related to sports, for example:
  - major league
  - sport team
- `OTHER`–Common nouns for organizations that do not fit into a more specific subtype, for example:
  - Palestinian and Lebanese organizations
  - largest opposition party

9.1.9 COMMON_PERSON

Common nouns for persons, for example:

- ceremony ministers
- injured members
- submarine crew
- 58-year-old man
- math teacher

9.1.10 COMMON_PRECURSOR

Common names for weapon precursors, extracted as one of the following subtypes:

- `CHEMICAL`
  - precursor chemical material
- `NUCLEAR`
  - precursor nuclear material

9.1.11 COMMON_REGION

Common nouns for different regions, extracted as one of the following subtypes:

`MAJOR`– Common nouns for states and provinces. For example:

- historical provinces
- home state

`MINOR`– Common nouns for the entirety of district areas. For example:

- millionaire counties
- development district

9.1.12 COMMON_VEHICLE

Methods of transportation, extracted as one of the following subtypes:

- `AIR`
  - dirigible
  - Black Hawk helicopter
  - jetliner
- `LAND`
  - car

  - motorbike
- `WATER`
  - cruiseliner
  - boat
  - Russian submarine
- `SUBAREA` (portions of vehicles where humans can fit)
  - cockpit
  - engine room
- `OTHER` (vehicles what do not fit into a more specific subtype)
  - chopper

9.1.13 COMMON_WEAPON

Common names for weapons, extracted as one of the following subtypes:

- `BIOLOGICAL`
  - attractive biological threat agent
- `CHEMICAL`
  - nerve agent
- `EXPLODING`
  - rocket-propelled grenades
- `NUCLEAR`
  - radioactive materials
- `PROJECTILE`
  - bullets
- `SHARP`
  - machetes
- `SHOOTING`
  - rifle
- `OTHER`
  - weapons

Index

A

advanced extraction support for
  language modules 31
advanced linguistic analysis support
  for language modules 16
advanced parsing
  English 119

B

basic-level extraction support for
  language modules 31
  Greek 180
  Hungarian 183
  Polish 252
  Romanian 265
  Thai 327
  Turkish 329
Bokmål
  case variants 230
  character encodings 229
  deaccented characters 230
  entity types 239
  expanded inflectional stemmer 230
  extraction 239
  grouping 238
  hyphenation 230
  linguistic processing 228
  noun groups 239
  part-of-speech tagging 232
  standard stemmer 230
  stemming 230
  typewriter forms of accented letters
    230
  word segmentation 229
Bokmål language module reference
  228

C

case normalization 15, 25, 26
  rules 20
case variante
  Italian 185
case variants
  Bokmål 230
  Catalan 46
  Croatian 79
  Czech 85
  Danish 92
case variants *(continued)*
  Dutch 102
  English 113
  French 137
  German 158
  Hungarian 182
  Nynorsk 242
  Portuguese 255
  Romanian 264
  Serbian 277
  Slovak 283
  Slovenian 290
  Spanish 300
Catalan
  case variants 46
  character encodings 43
  deaccented characters 46
  entity types 52
  expanded inflectional stemmer 46
  extraction 52
  grouping 51
  hyphenation 46
  linguistic processing 43
  noun groups 52
  part-of-speech tagging 47
  standard stemmer 44
  stemming 44
  unfound words 51
  word segmentation 44
Catalan language module reference
  43
CCJT languages 19, 20
character encodings for
  Bokmål 229
  Catalan 43
  Chinese - Simplified 53
  Chinese - Traditional 72
  Croatian 78
  Czech 84
  Danish 90
  Dutch 100
  English 110
  French 134
  German 154
  Greek 179
  Hungarian 181
  Italian 183
  Japanese 201
  Korean 216
  Nynorsk 240
  Polish 251
character encodings for *(continued)*
  Portuguese 253
  Romanian 263
  Russian 266
  Serbian 276
  Slovak 282
  Slovenian 289
  Spanish 297
  Swedish 316
  Thai 326
  Turkish 328
characters with missing diacritics
  Serbian 277
common mentions content
  rules 393
common noun mentions
  Chinese - Simplified 67
  English 41
  Simplified Chinese 41
complete tags 25
compound analysis 21, 24
  Chinese 74
  Dutch 103
  German 159
  Korean 219
  Swedish 319
compound stemmer
  Swedish 319
compound word stemming 24
compound words 16
Croatian
  case variants 79
  character encodings 78
  deaccented characters 79
  entity types 83
  expanded stemmer 79
  extraction 83
  grouping 82
  linguistic processing 78
  noun groups 83
  part-of-speech tagging 80
  standard stemmer 79
  stemming 79
Croatian language module reference
  78
customizing
  dictionaries 12, 29
  extraction 12, 29
  rules 12, 29
Czech
  case variants 85

Czech *(continued)*
character encodings 84
deaccented characters 85
entity types 89
expanded stemmer 85
extraction 89
grouping 89
linguistic processing 84
noun groups 89
part-of-speech tagging 86
standard stemmer 84
stemming 84
unfound words 88
word segmentation 84
Czech language module reference 83

D

Danish
case variants 92
character encodings 90
deaccented characters 92
entity types 99
expanded inflectional stemmer 92
extraction 99
grouping 98
hyphenation 92
linguistic processing 90
noun groups 99
part-of-speech tagging 93
standard stemmer 91
stemming 91
typewriter forms of accented letters 92
word segmentation 90
Danish language module reference 90
deaccented characters
Bokmål 230
Catalan 46
Czech 85
Danish 92
Dutch 102
French 137
German 158
Hungarian 182
Italian 185
Nynorsk 242
Portuguese 255
Romanian 264
Slovak 283
Slovenian 290
Spanish 300
Swedish 318
derivational stemmer
English 114
derivational stemming 21, 24
dictionaries 12, 29, 30, 31
document analysis 15, 26
Dutch
case variants 102
character encodings 100
compound analysis 103
deaccented characters 102
entity types 109
expanded inflectional stemmer 102
extraction 109
grouping 108
hyphenation 102
linguistic processing 100
noun groups 109
part-of-speech tagging 105
standard stemmer 101
stemming 101
word segmentation 100
Dutch language module reference 99

E

emoticon extraction examples
English 338
emoticons
extracting voice of the customer content 338
encoding identification 15
English
advanced parsing 119
case variants 113
character encodings 110
derivational stemmer 114
emoticon extraction examples 338
entity types 119
expanded inflectional stemmer 113
extraction 119
hyphenation 113
inflectional stemmer guesser 115
linguistic processing 110
noun groups 126
part-of-speech tagging 115
profanity extraction examples 342
public sector content rules 354
request extraction examples 340
sentiment extraction examples 334
standard stemmer 112
stemming 112
subtypes 119
unfound words 118
word segmentation 110
English language module reference 109
enterprise content
rules 343
entity
defined 11
named entities 33
subtype 30
entity extraction 29
entity types
Bokmål 239
Catalan 52
Chinese - Simplified 59
Chinese - Traditional 78
common noun mentions 41
Croatian 83
Czech 89
Danish 99
Dutch 109
English 119
French 143
German 168
Italian 192
Japanese 211
Korean 225
Nynorsk 250
Portuguese 263
public sector entities
Simplified Chinese 387
Russian 272
Serbian 281
Slovak 288
Slovenian 296
Spanish 306
Swedish 325
expanded inflectional stemmer 21
Bokmål 230
Catalan 46
Danish 92
Dutch 102
English 113
French 137
German 158
Italian 185
Nynorsk 242
Portuguese 255
Spanish 300
Swedish 318
expanded inflectional stemming 23
expanded stemmer
Chinese 55
Croatian 79
Czech 85
Hungarian 182
Japanese 207
Romanian 264
Serbian 277
Slovak 283
Slovenian 290

extracting
common mentions information 393
emoticons using voice of the customer content 338
enterprise information 343
profanities using voice of the customer content 342
public sector entities - Simplified Chinese 387
public sector information 353
requests using voice of the customer content 339
sentiments using voice of the customer content 333
voice of the customer information 331
extraction
Bokmål 239
Catalan 52
Chinese - Simplified 58
Chinese - Traditional 78
common mentions content 393
Croatian 83
customizing 12, 29
Czech 89
Danish 99
Dutch 109
English 119
enterprise content 343
French 142
German 168
Greek 180
Hungarian 183
Italian 192
Japanese 210
Korean 225
levels of support for language modules 31
Nynorsk 250
overview 11, 29
Polish 252
Portuguese 262
public sector content 353
resource files 30
Romanian 265
Russian 271
Serbian 281
Slovak 288
Slovenian 296
Spanish 305
subtypes 30
Swedish 325
Thai 327
Turkish 329
voice of the customer content 331
extraction rules 12, 29, 30, 31
common mentions content 393
enterprise content 343
public sector content 353
English 354
voice of the customer 331

F

fact
defined 11
fact extraction 29
features, in each language 16
French
case variants 137
character encodings 134
deaccented characters 137
entity types 143
expanded inflectional stemmer 137
extraction 142
hyphenation 137
inflectional stemmer guesser 139
linguistic processing 134
noun groups 148
part-of-speech tagging 139
request extraction examples 340
sentiment extraction examples 335
standard stemmer 135
stemming 135
subtypes 143
unfound words 142
word segmentation 134
French language module reference 134

G

German
case variants 158
character encodings 154
compound analysis 159
deaccented characters 158
entity types 168
expanded inflectional stemmer 158
extraction 168
hyphenation 158
inflectional stemmer guesser 159
linguistic processing 154
non-decompounding stemmer 162
noun groups 173
part-of-speech tagging 163
request extraction examples 341
sentiment extraction examples 336
standard stemmer 157
stemming 156
subtypes 168
German *(continued)*
unfound words 167
word segmentation 155
German language module reference 154
Greek
character encodings 179
extraction 180
linguistic processing 179
stemming 179
word segmentation 179
Greek language module reference 179
grouping
Bokmål 238
Catalan 51
Croatian 82
Czech 89
Danish 98
Dutch 108
Nynorsk 249
Portuguese 262
Serbian 281
Slovak 288
Slovenian 296
Swedish 324

H

Hungarian
case variants 182
character encodings 181
deaccented characters 182
expanded stemmer 182
extraction 183
linguistic processing 180
standard stemmer 181
stemming 181
word segmentation 181
Hungarian language module reference 180
hyphenation
Bokmål 230
Catalan 46
Danish 92
Dutch 102
English 113
French 137
German 158
Italian 185
Nynorsk 242
Portuguese 255
Spanish 300
Swedish 318

I

inflectional stemmer guesser 23
English 115
French 139
German 159
Italian 186
Spanish 301
inflectional stemming 16
Italian
case variants 185
character encodings 183
deaccented characters 185
entity types 192
expanded inflectional stemmer 185
extraction 192
hyphenation 185
inflectional stemmer guesser 186
linguistic processing 183
noun groups 196
part-of-speech tagging 187
standard stemmer 184
stemming 184
subtypes 192
word segmentation 183
Italian language module reference 183

J

Japanese
character encodings 201
entity types 211
expanded stemmer 207
extraction 210
linguistic processing 201
part-of-speech tagging 209
standard stemmer 206
stemming 206
subtypes 211
word segmentation 201
Japanese language module reference 201

K

Korean
character encodings 216
compound analysis 219
entity types 225
extraction 225
linguistic processing 216
noun-noun compounds 219
noun-verb compounds 219
part-of-speech tagging 220
standard stemmer 218
stemming 218
Korean *(continued)*
subtypes 225
unfound words 224
word segmentation 216
Korean language module reference 216

L

language feature matrix 16
language identification 15
language module
defined 11
language module reference
Bokmål 228
Catalan 43
Croatian 78
Czech 83
Danish 90
Dutch 99
English 109
French 134
German 154
Greek 179
Hungarian 180
Italian 183
Japanese 201
Korean 216
Nynorsk 240
Polish 251
Portuguese 252
Romanian 263
Russian 265
Serbian 275
Simplified Chinese 53
Slovak 282
Slovenian 289
Spanish 297
Swedish 316
Thai 325
Traditional Chinese 72
Turkish 327
language modules 30
levels of extraction support 31
levels of linguistic analysis support 16
language modules reference 43
linguistic analysis
levels of support for language modules 16
overview 11
linguistic analysis support
advanced 16
standard 16
linguistic processing
Bokmål 228
linguistic processing *(continued)*
Catalan 43
Chinese - Simplified 53
Chinese - Traditional 72
Croatian 78
Czech 84
Danish 90
Dutch 100
English 110
French 134
German 154
Greek 179
Hungarian 180
Italian 183
Japanese 201
Korean 216
Nynorsk 240
Polish 251
Portuguese 252
Romanian 263
Russian 266
Serbian 276
Slovak 282
Slovenian 289
Spanish 297
Swedish 316
Thai 326
Turkish 327

M

multiword units 16, 19

N

named entities 33
natural language processing (NLP) 11
features used 9
NLP
capabilities used 9
non-decompound stemming 24
non-decompounding stemmer
German 162
noun group
Bokmål 239
Catalan 52
Croatian 83
Czech 89
Danish 99
Dutch 109
English 126
French 148
German 173
Italian 196
Nynorsk 250
Portuguese 263

noun group *(continued)*
Serbian 281
Simplified Chinese 63
Slovak 288
Slovenian 296
Spanish 310
Swedish 325
Traditional Chinese 78
NOUN_GROUP 33
Nynorsk
case variants 242
character encodings 240
deaccented characters 242
entity types 250
expanded inflectional stemmer 242
extraction 250
grouping 249
hyphenation 242
linguistic processing 240
noun groups 250
part-of-speech tagging 243
standard stemmer 241
stemming 241
typewriter forms of accented letters 242
word segmentation 240
Nynorsk language module reference 240

P

part-of-speech tagging 15, 16, 26
Bokmål 232
Catalan 47
Chinese - Simplified 56
Chinese - Traditional 76
complete tags 25
Croatian 80
Czech 86
Danish 93
Dutch 105
English 115
French 139
German 163
Italian 187
Japanese 209
Korean 220
Nynorsk 243
Portuguese 256
Russian 267
Serbian 278
Slovak 284
Slovenian 291
Spanish 301
Swedish 320
tag name conventions 26
part-of-speech tagging *(continued)*
umbrella tags 25
Polish
character encodings 251
extraction 252
linguistic processing 251
stemming 252
word segmentation 251
Polish language module reference 251
Portuguese
case variants 255
character encodings 253
deaccented characters 255
entity types 263
expanded inflectional stemmer 255
extraction 262
grouping 262
hyphenation 255
linguistic processing 252
noun groups 263
part-of-speech tagging 256
standard stemmer 254
stemming 253
word segmentation 253
Portuguese language module reference 252
predefined entity type support 33
profanities
extracting voice of the customer content 342
profanity extraction examples
English 342
public sector content
entity types
Simplified Chinese 387
languages supported 353
rules
English 354
types of information extracted
English 353
Simplified Chinese 387
punctuation 20

R

request extraction examples
English 340
French 340
German 341
Spanish 341
requests
extracting with voice of the customer content 339
resource files, extraction
dictionaries 30
extraction rules 30
resource files, extraction *(continued)*
language modules 30
Romanian
case variants 264
character encodings 263
deaccented characters 264
expanded stemmer 264
extraction 265
linguistic processing 263
standard stemmer 264
stemming 264
word segmentation 264
Romanian language module reference 263
rules 30
common mentions content 393
enterprise content 343
for case normalization 20
public sector content 353
English 354
voice of the customer content 331
Russian
character encodings 266
entity types 272
extraction 271
linguistic processing 266
part-of-speech tagging 267
stemming 266
subtypes 271
word segmentation 266
Russian language module reference 265

S

segment generation 15, 19
sentiment extraction examples
English 334
French 335
German 336
Spanish 337
sentiments
extracting voice of the customer content 333
Serbian
case variants 277
character encodings 276
characters with missing diacritics 277
entity types 281
expanded stemmer 277
extraction 281
grouping 281
linguistic processing 276
noun groups 281
part-of-speech tagging 278

Serbian *(continued)*
standard stemmer 276
stemming 276
unfound words 281
word segmentation 276
Serbian language module reference 275
Simplified Chinese
character encodings 53
common noun mentions 67
entity types 59
public sector content 387
expanded stemmer 55
extraction 58
linguistic processing 53
noun groups 63
part-of-speech tagging 56
standard stemmer 54
stemming 54
subtypes 58
word segmentation 53
Simplified Chinese language module reference 53
Slovak
case variants 283
character encodings 282
deaccented characters 283
entity types 288
expanded stemmer 283
extraction 288
grouping 288
linguistic processing 282
noun groups 288
part-of-speech tagging 284
standard stemmer 283
stemming 283
unfound words 287
word segmentation 282
Slovak language module reference 282
Slovenian
case variants 290
character encodings 289
deaccented characters 290
entity types 296
expanded stemmer 290
extraction 296
grouping 296
linguistic processing 289
noun groups 296
part-of-speech tagging 291
standard stemmer 289
stemming 289
unfound words 295
word segmentation 289
Slovenian language module reference 289
Spanish
case variants 300
character encodings 297
deaccented characters 300
entity types 306
expanded inflectional stemmer 300
extraction 305
hyphenation 300
inflectional stemmer guesser 301
linguistic processing 297
noun groups 310
part-of-speech tagging 301
request extraction examples 341
sentiment extraction examples 337
standard stemmer 298
stemming 298
subtypes 305
unfound words 305
word segmentation 297
Spanish language module reference 297
specialized extraction
common mentions content 393
enterprise content 343
public sector content 353
voice of the customer content 331
standard extraction support for language modules 31
standard inflectional stemming 21, 22
standard linguistic analysis support for language modules 16
standard stemmer
Bokmål 230
Catalan 44
Chinese 54
Chinese - Traditional) 73
Croatian 79
Czech 84
Danish 91
Dutch 101
English 112
French 135
German 157
Hungarian 181
Italian 184
Japanese 206
Korean 218
Nynorsk 241
Portuguese 254
Romanian 264
Serbian 276
Slovak 283
Slovenian 289
Spanish 298
Swedish 317
stemming 15, 21
compound word stemming 24
derivational 24
expanded inflectional 23
inflectional stemmer guesser 23
non-decompounding 24
standard inflectional 22
tagged 26
unknown words 25
stemming in
Bokmål 230
Catalan 44
Chinese 54, 73
Croatian 79
Czech 84
Danish 91
Dutch 101
English 112
French 135
German 156
Greek 179
Hungarian 181
Italian 184
Japanese 206
Korean 218
Nynorsk 241
Polish 252
Portuguese 253
Romanian 264
Russian 266
Serbian 276
Slovak 283
Slovenian 289
Spanish 298
Swedish 317
Thai 326
Turkish 328
subtypes
defined 30
English 119
French 143
German 168
Italian 192
Japanese 211
Korean 225
Russian 271
Simplified Chinese 58
Spanish 305
Swedish
character encodings 316
compound stemmer 319
deaccented characters 318
entity types 325
expanded inflectional stemmer 318
extraction 325
grouping 324

Swedish *(continued)*
hyphenation 318
linguistic processing 316
noun groups 325
part-of-speech tagging 320
standard stemmer 317
stemming 317
typewriter forms of accented letters 318
word segmentation 317
Swedish language module reference 316

T

tag name conventions 26
tagged stemming 15, 16, 26
tagging, see part-of-speech tagging 25
Thai
character encodings 326
extraction 327
linguistic processing 326
stemming 326
word segmentation 326
Thai language module reference 325
Traditional Chinese
character encodings 72
entity types 78
expanded stemmer 74
extraction 78
linguistic processing 72
noun groups 78
part-of-speech tagging 76
standard stemmer 73
stemming 73
word segmentation 72
Traditional Chinese language module reference 72
Turkish
character encodings 328
extraction 329
linguistic processing 327
stemming 328
Turkish *(continued)*
word segmentation 328
Turkish language module reference 327
typewriter forms of accented letters
Bokmål 230
Danish 92
Nynorsk 242
Swedish 318

U

umbrella tags 25
unfound words 26
Catalan 51
Czech 88
English 118
French 142
German 167
Korean 224
Serbian 281
Slovak 287
Slovenian 295
Spanish 305
unknown words
stemming of 25

V

voice of the customer content
emoticon extraction examples
English 338
entity types 331
extracting emoticons 338
extracting profanities 342
extracting requests 339
extracting sentiments 333
languages supported 331
profanity t extraction examples
English 342
request extraction examples
English 340
French 340
voice of the customer content *(continued)*
request extraction examples *(continued)*
German 341
Spanish 341
sentiment extraction examples
English 334
French 335
German 336
Spanish 337

W

white space languages 19
word breaking 27
word segmentation 15, 16, 19
Bokmål 229
Catalan 44
Chinese 53, 72
Czech 84
Danish 90
Dutch 100
English 110
French 134
German 155
Greek 179
Hungarian 181
Italian 183
Japanese 201
Korean 216
Nynorsk 240
Polish 251
Portuguese 253
Romanian 264
Russian 266
Serbian 276
Slovak 282
Slovenian 289
Spanish 297
Swedish 317
Thai 326
Turkish 328