Data Category Registry: Morpho-syntactic and Syntactic Profiles

Gil Francopoulo, Thierry Declerck, Virach Sornlertlamvanich,

Eric de la Clergerie, Monica Monachini

affiliation of first author: Tagmatica, 126 rue de Picpus, 75012 Paris, France

gil.francopoulo@wanadoo.fr, declerck@dfki.de, virach@tcllab.org,
Eric.Clergerie@inria.fr, monica.monachini@ilc.cnr.it

Abstract

After a brief presentation of the data model, we describe a work in progress to define an initial set of morpho-syntactic and syntactic data categories dedicated to NLP applications. The aim is to improve interoperability among language resources and to optimize the process leading to their integration in applications. The main point is to be sure that when a language resource makes use of a value, the other language resources and programs have the same interpretation for this given value. From a practical point of view, these values are collected from existing lists, discussed, extended, and then recorded within a freely accessible data base: the ISO Data Category Registry.

1. Introduction

Data associated with language resources are identified and stored in a wide variety of environments like terminological data collections and NLP resources. With this respect, we believe that the production of a family of consensual ISO specifications and data can be a useful aid for the NLP actors.

In this paper, after a brief presentation of the data model, we describe a work in progress within ISO-TC37 whose aim is to gather and record data categories (Ide et al, 2004; Wright, 2004).

2. Context

The TC37 standards are currently elaborated as high level specifications and deal with word segmentation (ISO 24614), annotations (ISO 24611, 24612 and 24615), feature structures (ISO 24610), and lexicons (ISO 24613). These standards rely on low level specifications dedicated to constants, namely data categories (revision of ISO 12620), language codes (ISO 639), scripts codes (ISO 15924), country codes (ISO 3166) and Unicode (ISO 10646).

This bi-level approach will form a coherent family of standards with the following common and simple rules:

1) The high level specifications provide structural elements that are **decorated by the standardized constants**;
2) The low level specifications provide these standardized constants.

This decoupling is offered in order to provide a fine flexibility with regard to language and practice diversity. To be more concrete, for instance, in a high level structure such as a lexicon, different elements like a Lexical Entry and a Sense will be defined and linked together in order to allow the definition of different senses for a word, as follows:

```
<LexicalEntry>
  <feat att="partOfSpeech" val="noun"/>
    <Lemma>
      <feat att="writtenForm" val="bank"/>
    </Lemma>
  <Sense id="bank1">
    <Definition>
      <feat att="text" val="Business that keeps and lends money"/>
    </Definition>
  </Sense>
  <Sense id="bank2">
    <Definition>
      <feat att="text" val="Land along the side of a river"/>
    </Definition>
  </Sense>
</LexicalEntry>
```

In this example, LexicalEntry, Lemma, Sense, and Definition belong to high level specifications, more precisely: LMF. In contrast, partOfSpeech, noun, writtenForm, and text belong to low level specifications, more precisely: the Data Category Registry.

The usage of each of these high level elements is specified, together with their cardinality. The precise combination of high level elements and low level ones is not specified: this is left to the user. In other terms, the user selects the structural elements he needs, and provided that a suitable set of data categories is available, the user is able to decorate the structural elements for a given language.

3. Variations

For the high level specifications, a consensus must be found among what is to be considered as "the best practices" of our field. Implicitly, a mixed strategy based on "coherent union" of structures and a meta-model approach is often taken, depending on the agreement among the community.

The main criteria are:

- the various theoretical approaches;
- the languages covered;
- the type of resources (syntax, semantics …)

These three criteria apply on the data category side as well.

4. General objectives

The main objective of TC37 is interoperability and our work is done in the context of the revision of ISO-12620. The most frequently encountered problem is "how to merge data?" whereby the hardest sub-challenge is "how to compare data?".

To address these issues, first, the use of a uniform policy should contribute to system coherence and functionality. And secondly, each data category (DC) must be well defined in order to allow elementary operations like: "is DC-A the same notion as DC-B ?" "is DC-C more general (or more specific) than DC-D ?", or "is DC-E related somehow to DC-F ?".

5. Specific objectives

With this respect, we have two distinct objectives:

1) Test the current specification of the revision of ISO-12620 as a proof of concept ;

2) Concretely record an initial set of data for morpho-syntax and syntax.

The goal is not to create a rich network of links between data categories.

6. History of ISO-12620

The ISO standard 12620 was published in 1999. The document specifies the content of data categories and presents a long list of values, whose primary aim was be used in terminological data collections.

The revision of ISO-12620 is somehow different. The work started in 2003. The document is currently in Final Draft for International Standard (FDIS) stage[1], and the schedule is to reach International Standard (IS) publication in 2009. The development is twofold. The revised version specifies how the data categories will be described and managed, but in contrast to the initial version, the values will not be presented in the ISO document. The values will be managed within a database endorsed by ISO that is called the Data Category Registry (DCR).

Another point to mention is the type of high level structure that is addressed by the new set of data categories. The old version targeted only terminological data collections but the new version target is much broader. The coverage is all TC37 activities, which means that NLP applications are concerned, hence largely increasing the number of values. For instance, the old ISO-12620 had only three values for part of speech, namely: **noun**, **adjective** and **verb**, but now because of NLP data structures, values like **preposition** and **punctuation** are needed. So, instead of only three values, the list contains now one hundred values.

7. Current registry

As cited earlier, the 12620 revision work started in 2003, and a lot of energy has been spent along the years in various meetings and document writings, in order to find an operational consensus. The two tasks (DC specification and DC recording) were conducted in parallel with frequent interactions.

This model has been implemented in a system called "Syntax[2]" which is currently running and is located at http://syntax.inist.fr where about a dozen people have entered values, mainly in the domain of terminology, morpho-syntax, and syntax. The list of the current values is presented in Annex-B, with an indentation for the broader link information.

8. Data model

The current model allows a lot of options but we limit ourselves to a subset of features, as presented in the UML class diagram in Annex-A.

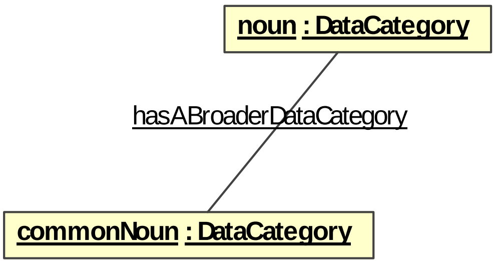
The registry is divided into profiles. A profile is a set of data categories. Each profile is associated with a team of experts with a convenor, who collectively represent a community of practice in the area of language resources. There are currently about ten profiles and as many or more sub-activities, such as terminology, metadata etc, covering all activities of ISO-TC-37. The current paper focuses on two profiles dedicated to NLP, namely the morpho-syntactic and syntactic profiles.

Many times, a data category belongs to only one profile, but a small number of them belongs to several profiles (e.g. part of speech).

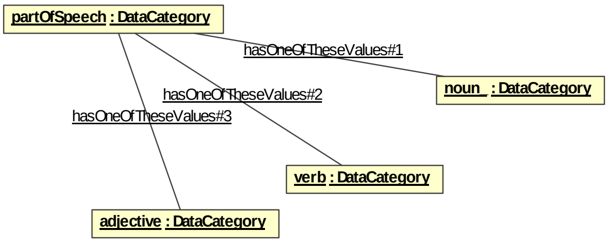
We differentiate between the notion of broader relation and the notion of value domain. The broader link allows a hierarchy of constants that forms an ontology. Example: a **common noun** is a more specialized value than **noun**.

[1] For a reader who is interested in reading the FDIS document, it may be accessed through the National Body channel: ASCII for US, DIN for Germany etc.

[2] The name is not very well chosen and does not mean that the system deals only with syntactic descriptions.



The notion of value domain is different. A value domain allows a set of valid values to be identified. In other terms, a value domain that is attached to a data category X provides a set of potential values for X and these values are themselves data categories. Example: **noun** is a value for **partOfSpeech**.



9. Data: methodology

We proceeded in three phases:

Phase-1: collating of candidates data categories
Phase-2: grouping, structuring, and redaction of a first draft of the definitions
Phase-3: revision

For the morpho-syntactic profile, a long initial list of data categories has been collected from:

- Current ISO-12620:1999
- Eagles and Multext-East
- Some values for Semitic languages coming from Sfax University

For the syntactic profile, an initial list was collected based on:

- Eagles
- Tiger (German project)
- Technolangue/Easy (French project)

Let us add that some values needed from TC37 standards like MAF (ISO-24611), SynAF (ISO-24615) (Declerck et al, 2006) and LMF (ISO-24613) (Francopoulo et al, 2006) have been added to the two profiles.

Each data category has an identifier that is English based. The name does not contain any spaces, and if more than one word is needed, it is expressed in so-called *camel case* (e.g. **commonNoun**) as specified in the revision of ISO-12620.

Currently each DC has a definition in English and French. Let us note that a lot of time has been devoted to write rigorous definitions, taking into account the various stable sources in our field. A definition may be complemented by a note.

A DC may be linked through a broader link to another DC. A DC may have a value domain.

Each DC has, at least, a name in English and one in French, which may be used directly for display without any transformation (e.g. **common noun**).

Currently, the ontology of values (through the broader link) is rather flat and does not exceed three levels. There are no constraints between DCs.

There is currently no indication concerning the use of a given DC for a specific language, but the new version will include a linguistic section that will enable some further constraints on value domains that may reflect specific usage in different object languages.

Thus, to reply to the question: "Is DC-A, the same notion as DC-B?", the user needs to compare identifier of DC-A to identifier of DC-B. If an explanation is needed to understand why two DCs are different, each DC has a precise definition for this purpose.

10. Data: organization

The number of values is rather huge, so in order to facilitate management, a series of directories[3] has been created within the two following profiles.

[3] A directory is equivalent to a sub-profile.

Morpho-syntactic profile:

Basics	61	items
These are general purpose linguistic constants, like: **comment**, **derivation**, **elision**, **foreignText**, and **label.**		
Cases	33	
Examples of values: **ablativeCase** or **dativeCase**.		
FormRelated	36	
These are constants for the specifications of forms like: **spokenForm**, **writtenForm**, **abbreviation**, **expansionVariation**, **transliteration**, **romanization**, **transcription**, **script.**		
Morphological Features excluding cases	82	
Attributes include for instance **grammaticalGender**, **mood** and **tense**. Values include, for instance, **feminine, indicative**, **present**.		
Operations	29	
Constants include for instance, **addAffix**, **addLemma**.		
Part of speech	120	
Part of speech values are structured with a top level set composed of 10 values like **noun** or **verb**. A very precise ontology is specified for grammatical words. Most of parts of speech are common to lexicons and annotations but two set of values (i.e. **punctuation** and **residual**) are specific to annotation and are not usually used in lexical descriptions[4].		
Register, dating and frequency	19	
Constants include, for instance, **slangRegister** or **rarelyUsed**.		
Total	380	items

In contrast to the values of the morpho-syntactic profile, which mainly concern the lexicon, most values in the syntactic profile deal with annotation.

Syntactic profile:

Basics	29	items
These are general purpose annotation constants, like: **tagging**, **standoffNotation**, **embeddedNotation**. A few of them like **negation** or **contiguous** concern lexicons.		
Constituency	27	
These comprise constants used to annotate constituency elements. Examples of values are: **chunk**, **declarativeClause**, **verbNucleus**, **nounPhrase**. Usual abbreviations like NP for **nounPhrase** are declared in the name section of the data category.		
Dependency	32	
These comprise constants used to annotate relation between syntactic elements. Examples of values are: **verbModifier**, **modifier**, **syntacticHead**, **subject, introducer, directObject**, **coordination**, **adjunct**. Let us note that a certain freedom is left to the user concerning the level of detail and the type of target: for instance, both **verbModifier** and **modifier** are proposed.		
Total	88	items

11. Problems encountered

As said earlier, we started from existing lists that are rather stable like those for Eagles or Multext-East. The problems that we encountered were that we had to write definitions. We searched in various sources and found some definitions that looked fine in isolation for some data categories, but they did not constitute a coherent set of definitions.

Linguistics is not a field with a common agreement on basic terms. As a matter of example, the entry "morphology" in Wikipedia, gives us a good view of these divergences. In linguistics, terms like "paradigm", "collocation", "morpheme", "ergative" have so many interpretations in the different theories that they are almost impossible to use in a normative context where a precise meaning is required.

Another problem we faced was that we had to write definitions that are valid for lexicons and annotation, and an important term like "word" does not have the same meaning in both contexts. A word in a lexicon is lexical entry that is associated with a lemma. A word in an annotation is an occurrence of an inflected form (in

[4] For the people working in terminology and lexicons, punctuation is usually not considered as a part of speech. The situation is rather different when the objective is to represent text specific structures like coordination in the context of syntactic annotation, in this case, a punctuation mark is usually considered as a plain word, and as such, needs a part of speech tagging.

an inflected language). Theses notions are rather different.

To deal with this problem, we carefully avoided dangerous terms and we delimited a secure set of terms. When needed, we formed multi-word expressions from secure components. This is the strategy that has been adopted in the DCR and in general within the ISO-TC37 family of standards.

12. Forthcoming data

The current database records values for West/East European languages and, to a certain extent, for Semitic languages. The rationale for such a strategy is that, first, it was easier for us to begin by these values because stable lists already existed for these languages. Secondly, we faced a "chicken and egg" situation: we rely on ISO voluntaries and no one will describe minority languages if the well-known languages were not covered.

We know that it is clearly not enough

Two other parallel tasks are currently being conducted. One task deals with Asian values within the NEDO project (Takenobu et al, 2006; Charoenporn et al, 2007; Shirai et al, 2008). A small set of values has been entered in the database. The other task deals with African values, and a study is being conducted by the ISO South African delegation, but the values have not been entered yet in the database.

Each value is associated with a version number to allow a stable compliance in case of modification. The rules for management and usage are defined in the ISO-12620 revision.

13. Forthcoming registry

The current system is rather simple. It permits to make simple interactive queries, to download the result of a query, to download a data category, a directory or a profile. The available formats are XML and HTML.

The registry has been populated with numerous data categories, but different users (including ourselves) asked for an upgrade with improved interface features and fully developed functionalities.

An improved model is currently being designed (2007-2008) in order to address two important issues namely the distinction between the language section (working language) and linguistic section (object language) and the ability to record constraints and richer relations. Another difference is that the relation "broader" has been renamed into "IsA".

The new model will be implemented in a system called "ISOcat" at http://www.isocat.org. This new system is currently in beta version and will be presented during LREC-2008 and described in (Kemps-Snijders et al, 2008; Wittenburg et al, 2007).

Instead of being based on traditional synchronized PHP programs, the new software is based on Java/Ajax technologies and promises to be more user friendly. The operational switch from Syntax to ISOcat is scheduled for the end of 2008.

14. Conclusion

The registry is far from being complete but it begins to be used within different ISO-TC37 based standard applications in order to be tested. The idea is to progressively increase the number and coverage of these data categories. The ambition is that the registry will become the reference point when using linguistic terms and data elements in lexicons and annotations in NLP context.

15. Acknowledgements

The work presented here is partially funded by the EU eContent-22236 LIRICS project and in part by the French ANR-Passage project (Action ANR-06 MDCA-013).

16. References

Charoenporn T., Thoongsup S., Sornlertlamvanish V., Isahara H. (2007) Thai Lexicon. SEALS Conference, Univ of Maryland, College Park. US

Declerck T. (2006) SynAF: Towards a standard for syntactic annotation. LREC Genoa.

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. (2006) Lexical Markup Framework (LMF). LREC Genoa.

Ide N., Romary L (2004) A Registry of Standard Data Categories for linguistic Annotation. LREC Lisboa.

ISO-12620:1999, Computer application in terminology - Data categories, ISO Geneva

Kemps-Snijders M., Windhouwer M., Wittenburg P., Wright S.E. (2008, forthcoming) A revised Data Model for the ISO Data Category Registry, submitted to TKE-2008, Copenhagen.

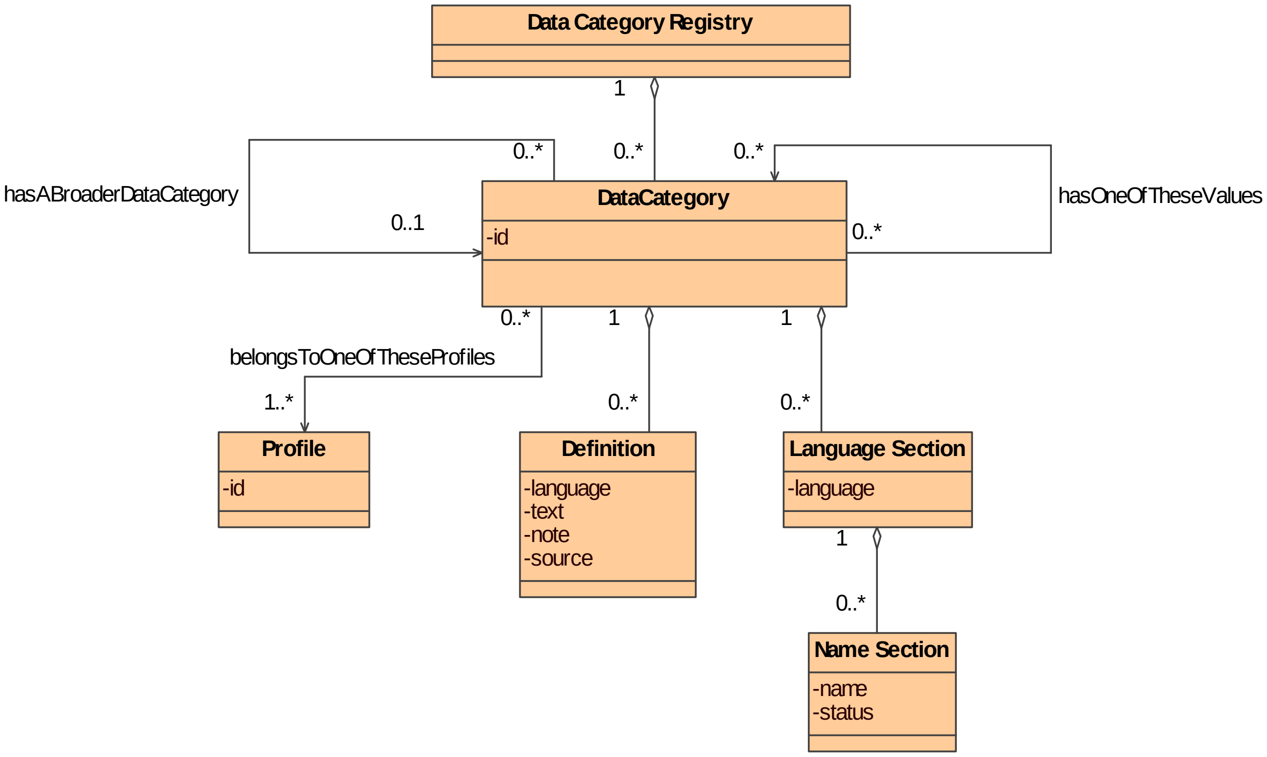
Shirai K., Tokunaga T., Huang CR., Hsieh SK, Kuo TY., Sornlertlamvanich, Charoenporn T. (2008) Constructing Taxonomy of Numerative Classifiers for Asian Languages IJCNLP Hyderabad, India

Takenobu T., Sornlertlamvanich V., Charoenporn T., Calzolari N., Monachini M., Soria C., Huang CR., Hao Y., Prevot L., Kiyoaki S. (2006) Infrastructure for standardization of Asian language resources COLING/ACL Sydney, Australia

Wittenburg P., Wright S.E. (2007) Infrastructure note on registry databases: technical note at http://www.tc37sc4.org/new_doc/iso_tc37_sc4_N436_ontology_memo_peter_Sue_busan2007.pdf

Wright S.E. (2004) A global data category registry for interoperable language resources: technical note at http://www.tc37sc4.org/new_doc/ISO_TC_37-4_N175_SEW-A_Global_Data_Category_Registry.pdf

Annex-A: UML class diagram of the portions of the current registry that we use



Annex-B: current set of values

Morpho-syntax: Basics

- agreement
- any
- approximate
- be
- coding
  - characterCoding
  - countryCoding
  - dateCoding
  - languageCoding
  - scriptCoding
- comment
- creationDate
- definition
- direction
- domain
- exact
- example
- expletive
- externalReference
- externalSystem
- have
- id
- image
- impossible
- label
- language
- leftEnvironment
- lexeme
- logicalOperator
  - logicalAnd
  - logicalNot
  - logicalOr
- logicalValue
  - no
  - yes
- macron
- namedEntity
- numValue
- pluralType
- position
- possible
- quotative
- rank
- reduplicationFunction
- reduplicationType
- required
- restriction
- rightEnvironment
- scope
- sound
- source
- space
- stringValue
- text
- type
- unspecified
- utterance
- value
- variation
- view
- word

Morpho-syntax: Cases

- case
  - abessiveCase
  - ablativeCase
  - absolutiveCase
  - accusativeCase
  - adessiveCase
  - aditiveCase
  - allativeCase
  - benefactiveCase
  - causativeCase
  - comitativeCase
  - dativeCase
  - delativeCase
  - elativeCase
  - equativeCase
  - ergativeCase
  - essiveCase
  - genitiveCase
  - illativeCase
  - inessiveCase
  - instrumentalCase
  - lativeCase
  - locativeCase
  - nominativeCase
  - obliqueCase
  - partitiveCase
  - prolativeCase
  - sociativeCase
  - sublativeCase
  - superessiveCase
  - terminativeCase
  - translativeCase
  - vocativeCase

Morpho-syntax: Form Related

- affix
  - infix
  - prefix
  - suffix
- affixRank
- allomorph
- apocope
- componentRank
- conjugated
- contextualVariation
- expansionVariation
- geographicalVariant
- graphicalSeparator
- homograph
- homonym
- homophone
- lemma
- lexicalType
- morpheme
  - etymologicalRoot
- native
- orthographyName
- patternType
- phoneticForm
- phoneticSeparator
- pinyin
  - nonSpacedPinyin
  - spacedPinyinAndTone
- reduplication
- root
- script
- stem
- stemRank
- symbol
- token
- writtenForm

Morpho-syntax: Morphological Features Excluding Cases

- activeVoice
- animate
- aorist
- bound
- cessative
- collective
- commonGender
- comparative
- conditional
- definite
- dual
- elInclusion
- elative
- feminine
- finite
- firstPerson
- fullArticle
- future
- gerundive
- honorific
- imperative
- imperfect
- imperfective
- inanimate
- inchoative
- indefinite
- indicative
- indifferent
- infinitive
- intensity
- masculine
- masdar
- middleVoice
- morphologicalFeature
  - animacy
  - aspect
  - cliticness
  - definiteness
  - degree
  - finiteness
  - grammaticalGender

  - grammaticalNumber
  - grammaticalTense
  - modificationType
  - negative
  - ownedNumber
  - ownerGender
  - ownerNumber
  - ownerPerson
  - person
    - objectPerson
    - subjectPerson
  - syntacticType
  - verbFormMood
  - voice
  - zuInclusion
- neuter
- nonFinite
- otherAnimacy
- participle
- passiveVoice
- past
- paucal
- perfective
- personal
- plural
  - brokenPlural
- positive
- possessive
- postModifier
- preModifier
- present
- quadrial
- referentType
- secondPerson
- shortArticle
- singular
- subjunctive
- superlative
- thirdPerson
- trial
- unaccomplished

Morpho-syntax: Operations

- abbreviation
- elision
- location
- operation
  - add
  - addAffix
  - addAfter
  - addBefore
  - addComponentLemma
  - addComponentStem
  - addFirstConsonant
  - addFirstVowel
  - addLemma
  - addLowerCaseComponentLemma
  - copy
  - derivation
  - remove
  - removeAfter
  - removeBefore
  - substitute
- operator
  - graphicalOperator
  - phoneticOperator
- romanization
- rule
- scheme
- transcription
- transformType
- transliteration

Morpho-syntax: Part of speech

- adjective
  - ordinalAdjective
  - participleAdjective
    - pastParticipleAdjective
    - presentParticipleAdjective
  - qualifierAdjective
- adposition
  - circumposition
  - postposition
  - preposition
    - compoundPreposition
    - fusedPreposition
    - simplePreposition
- adverb
  - generalAdverb
  - particleAdverb
- classifier
- conjunction
  - coordinatingConjunction
  - subordinatingConjunction
- determiner
  - article
    - definiteArticle
    - indefiniteArticle
    - partitiveArticle
  - demonstrativeDeterminer
  - exclamativeDeterminer
  - indefiniteDeterminer
  - interrogativeDeterminer
  - possessiveDeterminer
  - reflexiveDeterminer
  - relativeDeterminer
- interjection
- noun
  - commonNoun
  - countableNoun
  - diminutiveNoun
  - massNoun
  - properNoun
- numeral
  - numeralApprox
  - numeralBoth
  - numeralDigit
  - numeralLetter
  - numeralMForm
  - numeralRoman
- partOfSpeech
- particle
  - affirmativeParticle
  - comparativeParticle
  - conditionalParticle
  - coordinationParticle
  - distinctiveParticle
  - futureParticle
  - infinitiveParticle
  - interrogativeParticle
  - modalParticle
  - negativeParticle
  - possessiveParticle
  - relativeParticle
  - superlativeParticle
  - unclassifiedParticle
- pronoun
  - affixedPersonalPronoun
  - allusivePronoun
  - conditionalPronoun
  - demonstrativePronoun
  - emphaticPronoun
  - exclamativePronoun
  - impersonalPronoun
  - indefinitePronoun
  - interrogativePronoun
  - negativePronoun
  - personalPronoun
    - strongPersonalPronoun
    - weakPersonalPronoun
  - possessivePronoun
  - reciprocalPronoun
  - reflexivePronoun
  - relativePronoun
- punctuation
  - closePunctuation
    - closeBracket
    - closeCurlyBracket
    - closeParenthesis
  - mainPunctuation
    - declarativePunctuation
      - exclamativePoint
      - point
      - semiColon
      - suspensionPoints
    - interrogativePunctuation
      - questionMark
      - invertedQuestionMark
  - openPunctuation
    - openBracket
    - openCurlyBracket
    - openParenthesis
  - secondaryPunctuation
    - bullet
    - colon
    - comma
    - hyphen
    - invertedComma
    - quote

        - slash
    - unclassifiedPunctuation
- relationNoun
- residual
    - foreignText
    - foreignWord
    - formula
    - letter
    - unclassifiedResidual
- verb
    - auxiliary
    - copula
    - mainVerb
    - modal
- voiceNoun

Morpho-syntax: Register Dating Frequency

- benchLevelRegister
- commonlyUsed
- dating
- dialectRegister
- facetiousRegister
- formalRegister
- frequency
- inHouseRegister
- infrequentlyUsed
- ironicRegister
- modern
- neutralRegister
- old
- rarelyUsed
- register
- slangRegister
- tabooRegister
- technicalRegister
- vulgarRegister

Syntax: Basics

- annotation
    - morphosyntacticAnnotation
    - syntacticAnnotation
- annotationDeepness
- annotationStyle
- annotationType
- clitic
    - enclitic
    - proclitic
- constituency
- constituencyAndDependency
- contiguous
- deepParsing
- dependency
- doubleNegation
- embeddedNotation
- first
- mixedNotation
- negation
- next
- predicate
- previous
- propagation
- shallowParsing
- standoffNotation
- syntacticFeature
- tagging
- whType
- yesNoType

Syntax: Constituency

- grammaticalUnit
    - chunk
        - adjectiveChunk
        - adpositionChunk
        - adverbChunk
        - nounChunk
        - postpositionChunk
        - prepositionChunk
        - verbNucleus
    - clause
        - declarativeClause
        - imperativeClause
        - interrogativeClause
        - relativeClause
    - phrase
        - adjectivePhrase
        - adpositionPhrase
        - adverbPhrase
        - comparativePhrase
        - coordinatedPhrase
        - nounPhrase
        - postpositionPhrase
        - prepositionPhrase
        - prepositionVerbPhrase
        - superlativePhrase
        - verbPhrase
    - sentence

Syntax: Dependency

- adjunct
- apposed
- apposition
- attribute
- auxiliary
- complementizer
- coordination
- coordinator
- directObject
- function
- head
- introducer
- juxtaposition
- leftCoordinated
- modifier
    - adverbModifier
    - nounModifier
    - postnominalModifier
    - prenominalModifier
    - prepositionModifier
    - verbModifier
- relation
    - comparativeRelation
    - genitive
    - relativeRelation
    - superlativeRelation
- rightCoordinated
- subject
- syntacticArgument
- syntacticHead
- verbComplement