OWL/DL formalization of the MULTEXT-East morphosyntactic specifications

Christian Chiarcos
University of Potsdam, Germany
chiarcos@uni-potsdam.de

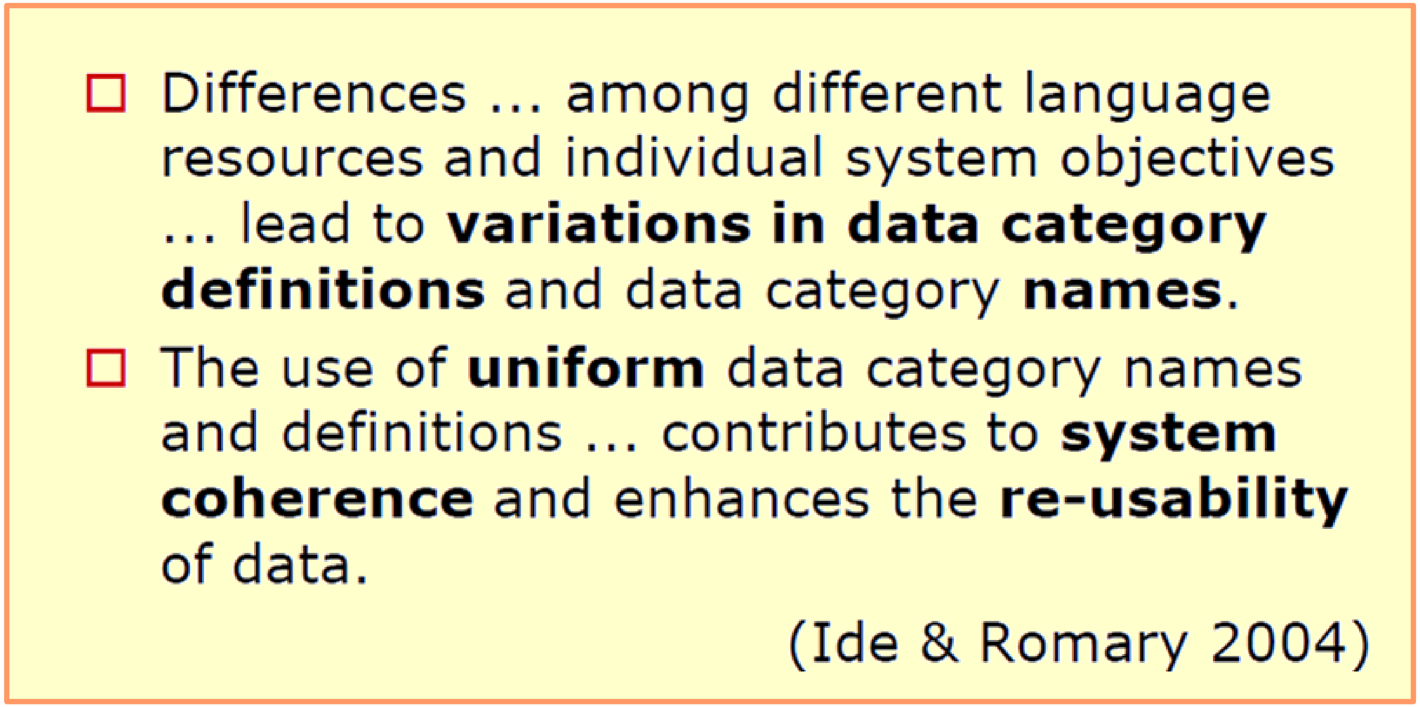
Tomaž Erjavec
Jožef Stefan Institute, Slovenia
tomaz.erjavec@ijs.si

5th Linguistic Annotation Workshop (LAW-V), held in conjunction with ACL-HLT 2011, Portland, Oregon, Jun 23rd-24th, 2011

OWL/DL formalization ...

- Background: Interoperability
- Multext-East (MTE) morphosyntactic specifications
- Building the MTE ontology
- Using the MTE ontology
- Revising the MTE ontology

Interoperability
The challenge



- □ Differences … among different language resources and individual system objectives … lead to **variations in data category definitions** and data category **names**.
- □ The use of **uniform** data category names and definitions … contributes to **system coherence** and enhances the **re-usability** of data.

(Ide & Romary 2004)

Interoperability Approaches

- Generalization and standardization
  - Multilingual tagset with categories, attributes and attribute values
    - EAGLES recommendations (Wilson & Leech 1996)
    - Multext-East (Dimitrova et al. 1998, Erjavec 2010)
  - Underspecified with respect to language-specific phenomena

Interoperability Approaches

- Generalization and standardization
- Centralization: data category registries
  - Central registry, may be extended by users
  - e.g. ISOcat (Kemps-Snijders et al. 2009)
  - Problems
    - There may be duplicates
      - e.g., vocative case (DC-1412, DC-2727, DC-3550)

    => Formalize relationships between data categories

Interoperability Approaches

- Generalization and standardization
- Centralization
- Formalization with ontologies
  - General Ontology of Linguistic Description (GOLD)

    (Farrar & Langendoen, 2003)
  - Concept taxonomy, relations, consistency constraints

Interoperability Approaches

(a) Generalization and standardization

(b) Centralization

(c) Formalization

Here

- Transformation of an existing resource of type (a) to one of type (c)
- Discussion of differences and benefits

MULTEXT-East (MTE) (http://nl.ijs.si/ME/V4)

- multilingual dataset for language engineering research and development (Dimitrova et al. 1998, Erjavec 2010)
  - Morphosyntactic specifications
  - Lexicons
  - Corpora
  - 16 languages (with morphosyntactic specifications)
    - Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Resian, Romanian, Russian, Serbian, Slovak, Slovene, Ukrainian

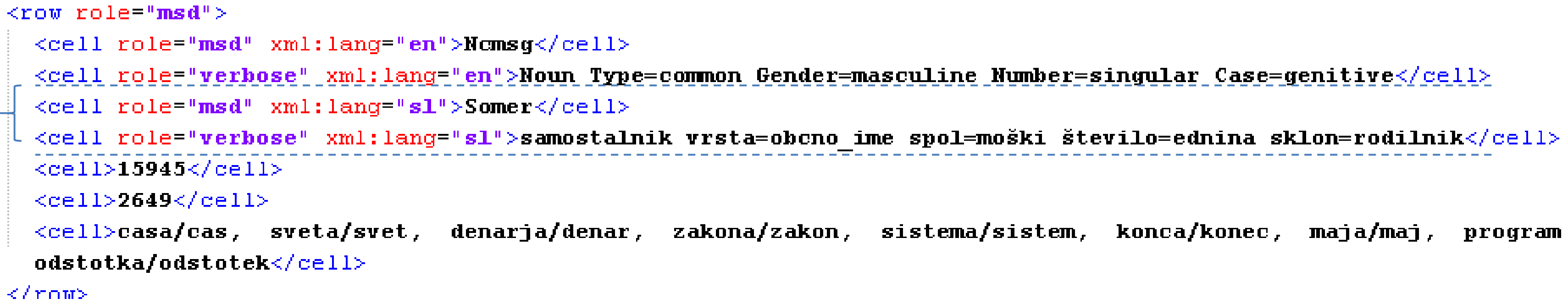
MULTEXT-East (MTE) Morphosyntactic specifications

- Positional tagset
  - Ncmsg

N	Noun	category
c	Type=common	attributes (and attribute values)
m	Gender=masculine	
s	Number=singular	
n	Case=nominative	

MULTEXT-East (MTE)
Language-specific specifications

- TEI XML document, defines tables where tags are explained

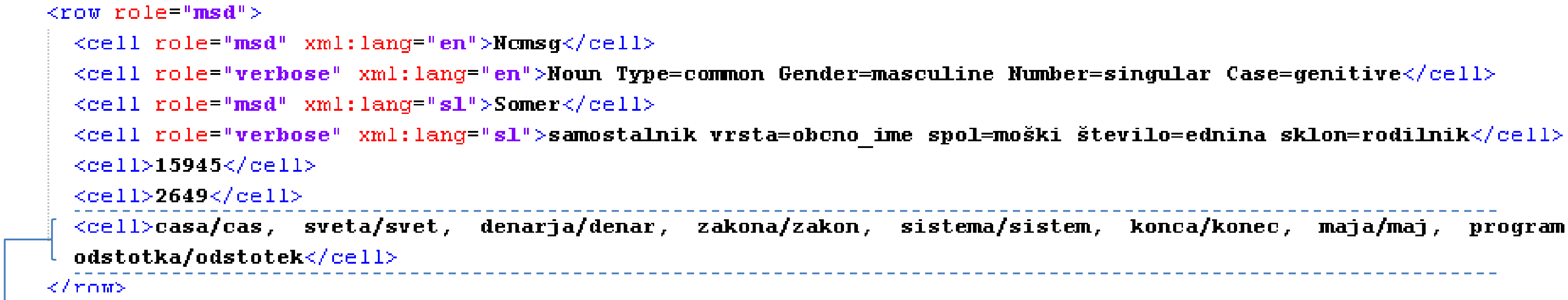


```
<row role="msd">
  <cell role="msd" xml:lang="en">Ncmsg</cell>
  <cell role="verbose" xml:lang="en">Noun Type=common Gender=masculine Number=singular Case=genitive</cell>
  <cell role="msd" xml:lang="sl">Somer</cell>
  <cell role="verbose" xml:lang="sl">samostalnik vrsta=obcno_ime spol=moški število=ednina sklon=rodilnik</cell>
  <cell>15945</cell>
  <cell>2649</cell>
  <cell>casa/cas, sveta/svet, denarja/denar, zakona/zakon, sistema/sistem, konca/konec, maja/maj, program
  odstotka/odstotek</cell>
</row>
```

Slene tag
Somer
and features

MULTEXT-East (MTE)
Language-specific specifications

- TEI XML document, defines tables where tags are explained



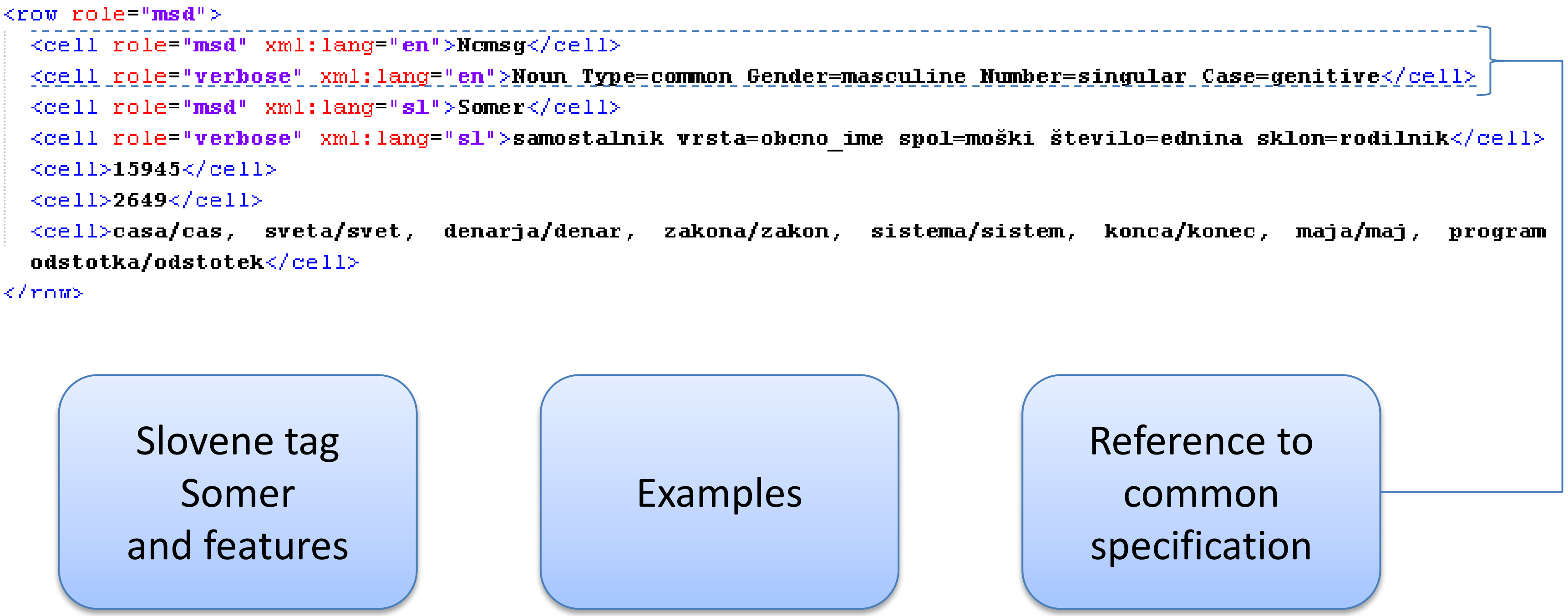
```
<row role="msd">
  <cell role="msd" xml:lang="en">Ncmsg</cell>
  <cell role="verbose" xml:lang="en">Noun Type=common Gender=masculine Number=singular Case=genitive</cell>
  <cell role="msd" xml:lang="sl">Somer</cell>
  <cell role="verbose" xml:lang="sl">samostalnik vrsta=obcno_ime spol=moški število=ednina sklon=rodilnik</cell>
  <cell>15945</cell>
  <cell>2649</cell>
  <cell>casa/cas,  sveta/svet,  denarja/denar,  zakona/zakon,  sistema/sistem,  konca/konec,  maja/maj,  program
  odstotka/odstotek</cell>
</row>
```

Slovene tag
Somer
and features

Examples

MULTEXT-East (MTE)
Language-specific specifications

- TEI XML document, defines tables where tags are explained



```
<row role="msd">
  <cell role="msd" xml:lang="en">Ncmsg</cell>
  <cell role="verbose" xml:lang="en">Noun Type=common Gender=masculine Number=singular Case=genitive</cell>
  <cell role="msd" xml:lang="sl">Somer</cell>
  <cell role="verbose" xml:lang="sl">samostalnik vrsta=obcno_ime spol=moški število=ednina sklon=rodilnik</cell>
  <cell>15945</cell>
  <cell>2649</cell>
  <cell>casa/cas, sveta/svet, denarja/denar, zakona/zakon, sistema/sistem, konca/konec, maja/maj, program
  odstotka/odstotek</cell>
</row>
```

Slovene tag
Somer
and features

Examples

Reference to
common
specification

MULTEXT-East (MTE) Common specifications

- TEI XML document
- Defines tables for categories

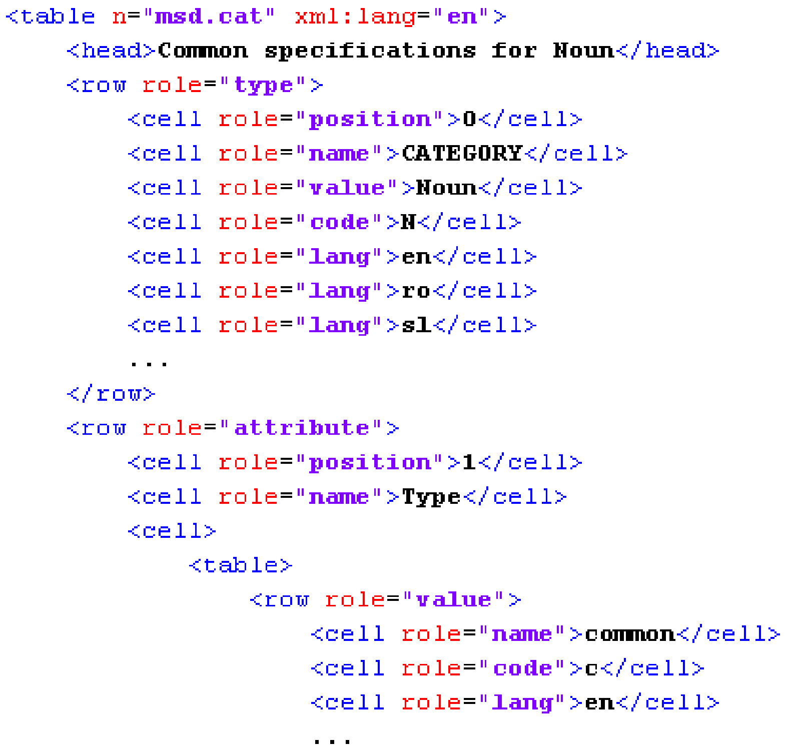
  //table/row[@role=‚type‘]

  attributes

  //table/row[@role=‚attribute‘]

  and attribute values

  //table/row[@role=‚value‘]



```
<table n="msd.cat" xml:lang="en">
    <head>Common specifications for Noun</head>
    <row role="type">
        <cell role="position">0</cell>
        <cell role="name">CATEGORY</cell>
        <cell role="value">Noun</cell>
        <cell role="code">N</cell>
        <cell role="lang">en</cell>
        <cell role="lang">ro</cell>
        <cell role="lang">sl</cell>
        ...
    </row>
    <row role="attribute">
        <cell role="position">1</cell>
        <cell role="name">Type</cell>
        <cell>
            <table>
                <row role="value">
                    <cell role="name">common</cell>
                    <cell role="code">c</cell>
                    <cell role="lang">en</cell>
                    ...
```

Building the MTE ontology

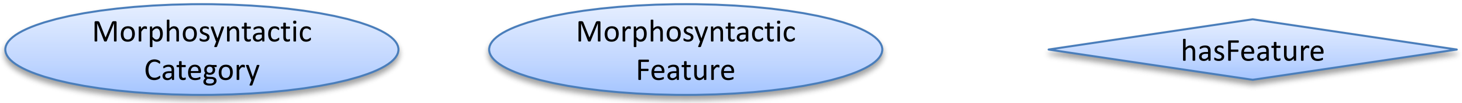
- OWL/DL
  - OWL: Web Ontology Language
    - RDF-based formalism to represent ontologies
      - Classes (concepts), instances (individuals), properties (relations)
  - DL: Description Logic
    - Decidable fragment of First Order Predicate Logic (FOPL)
      - join, intersection, complement
      - axioms: constraints on relations
    - Validation and inference

Building the MTE ontology Common specifications

1. Top-level concepts and properties

   - `mte:MorphosyntacticCategory,`
     `mte:MorphosyntacticFeature`
   - `mte:hasFeature :`

     `mte:MorphosyntacticCategory`
     `mte:MorphosyntacticFeature`



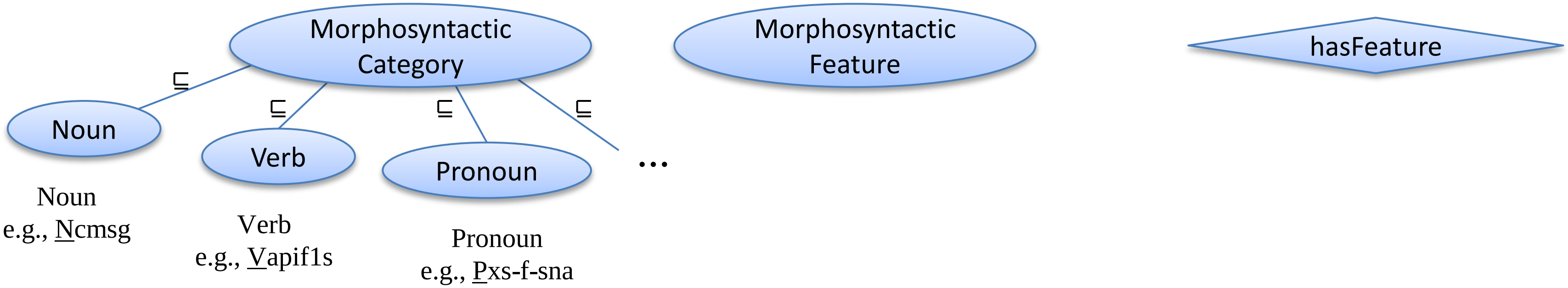
Building the MTE ontology
Common specifications

2. Direct children of `mte:MorphosyntacticCategory`

- For all MTE categories

  (POS tags in narrow sense, 1st position, e.g., Ncmsg)

  `mte:Noun`, `mte:Verb`, `mte:Pronoun`, ...



Morphosyntactic Category

Morphosyntactic Feature

hasFeature

⊑

Noun

Verb

Pronoun

...

Noun
e.g., Ncmsg

Verb
e.g., Vapif1s

Pronoun
e.g., Pxs-f-sna

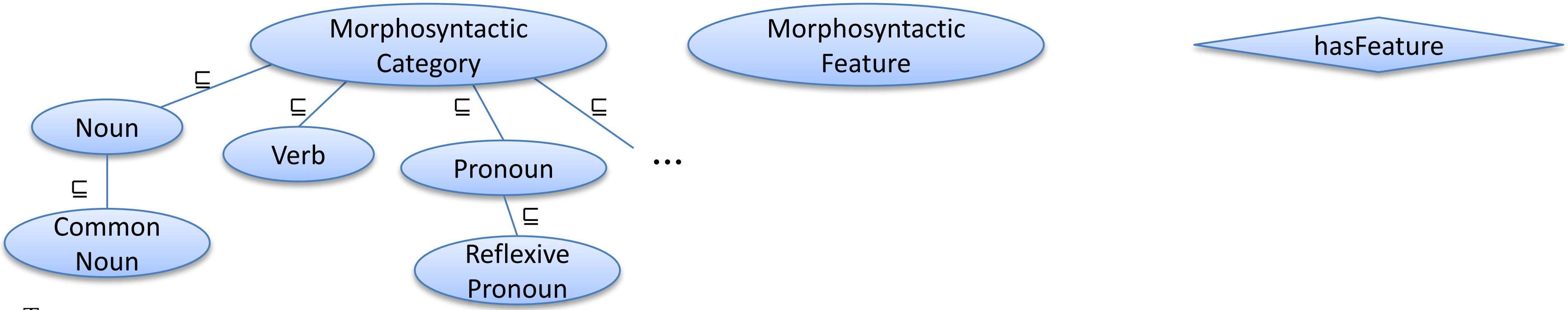
⊑ owl:subClassOf

Building the MTE ontology
Common specifications

3. Grandchildren of `mte:MorphosyntacticCategory`

– For MTE attribute Type

(2nd position, e.g., Ncmsg, Pxs-f-sna)



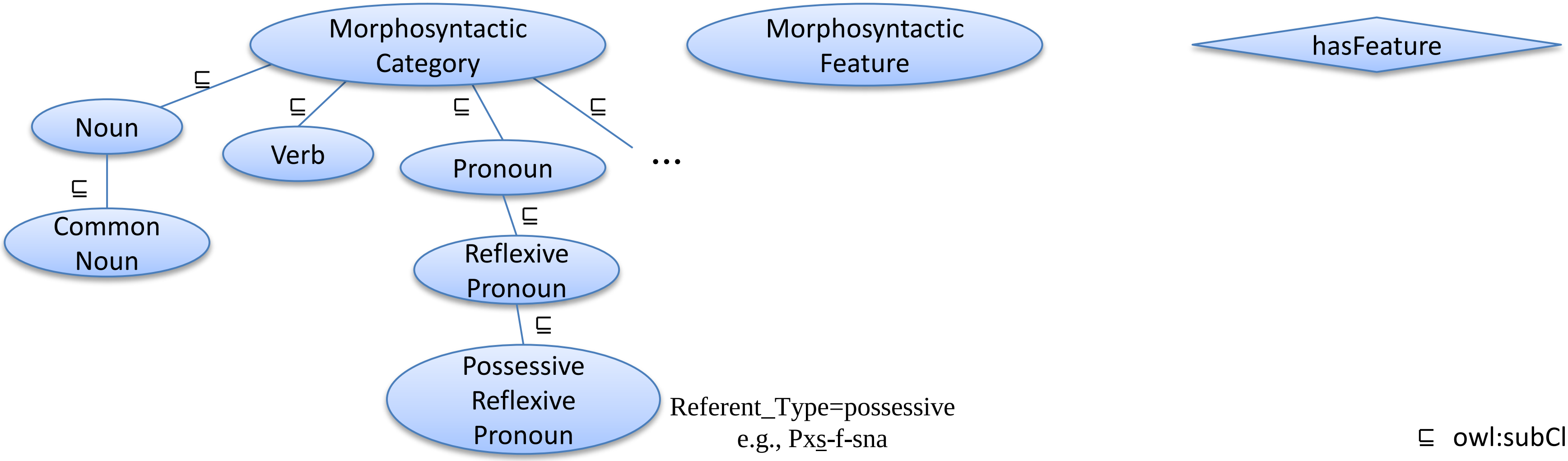
Type=common
e.g., Ncmsg

Type=reflexive
e.g., Pxs-f-sna

⊑ owl:subClassOf

Building the MTE ontology Common specifications

4. Great-grandchildren of `MorphosyntacticCategory`
   - For other MSD Type attributes (Wh_Type, Coord_Type, Sub_Type, Referent_Type)



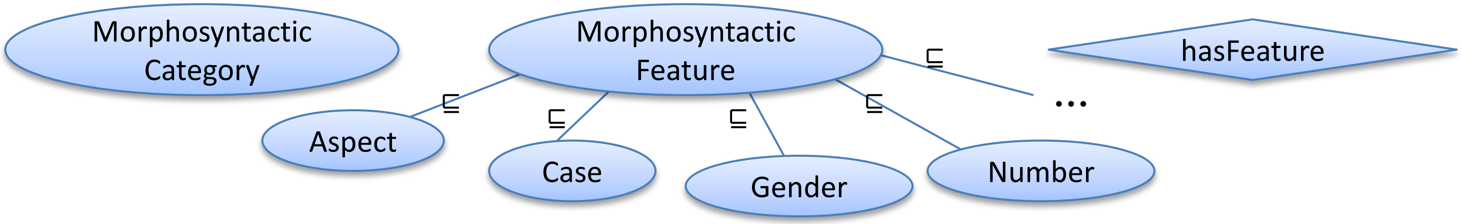
Referent_Type=possessive
e.g., Px<u>s</u>-f-sna

⊑ owl:subClassOf

Building the MTE ontology
Common specifications

5. All remaining attributes defined as children of `MorphosyntacticFeature`

e.g., `mte:Aspect`, `mte:Case`, ...

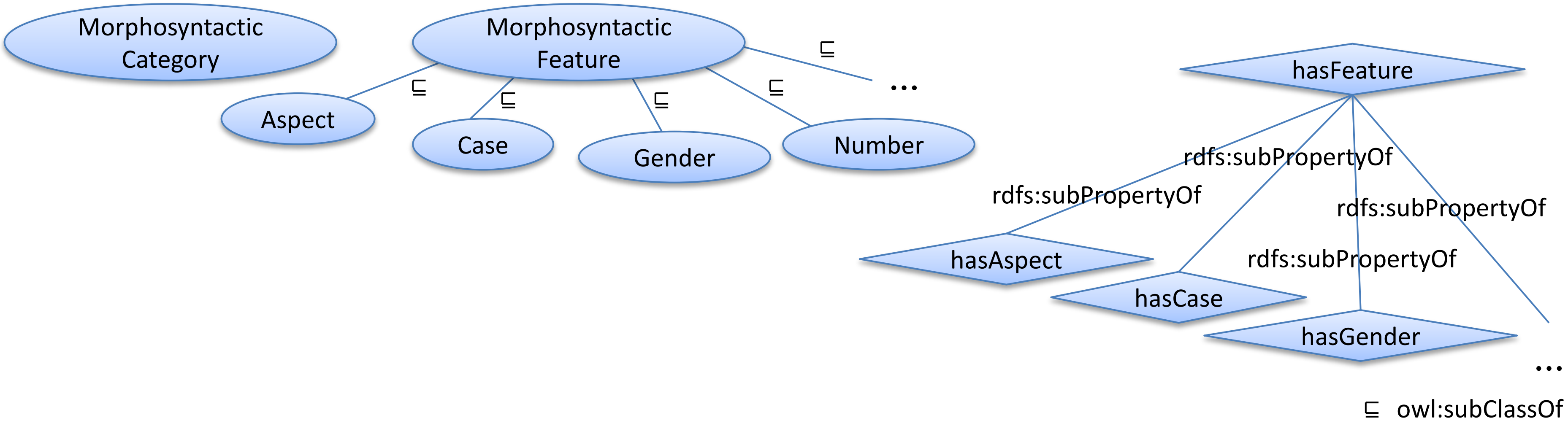


⊑ owl:subClassOf

Building the MTE ontology
Common specifications

5. All remaining attributes defined as children of `MorphosyntacticFeature`

6. … and a corresponding property is created

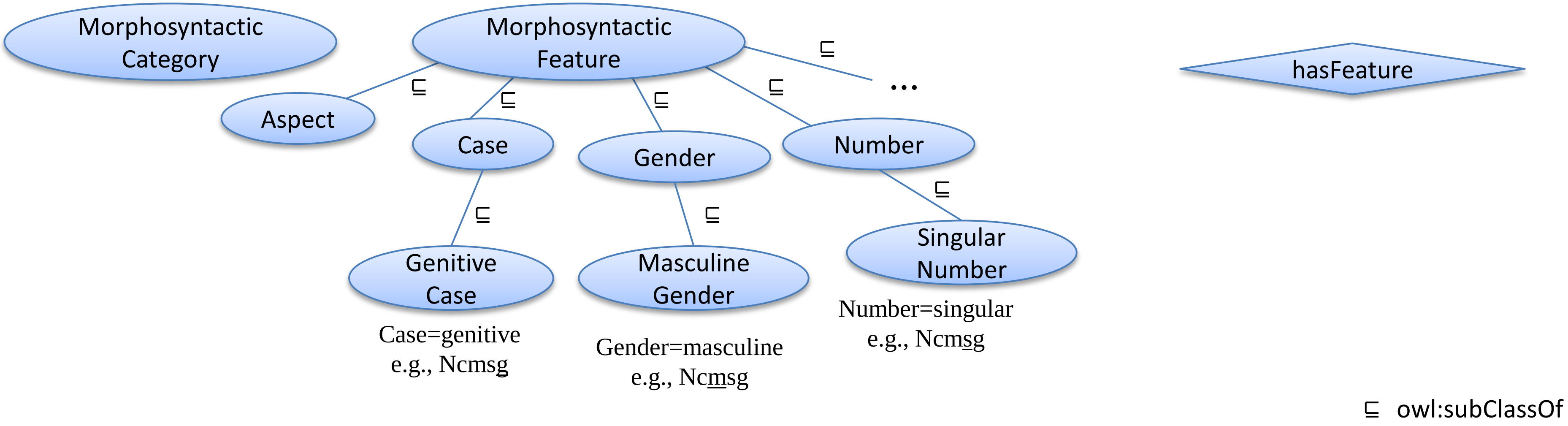


Building the MTE ontology Common specifications

7. All attribute values as subclasses of the corresponding `MorphosyntacticFeature`

e.g., Case=genitive =>

```
mte:GenitiveCase owl:subClassOf mte:Case
```

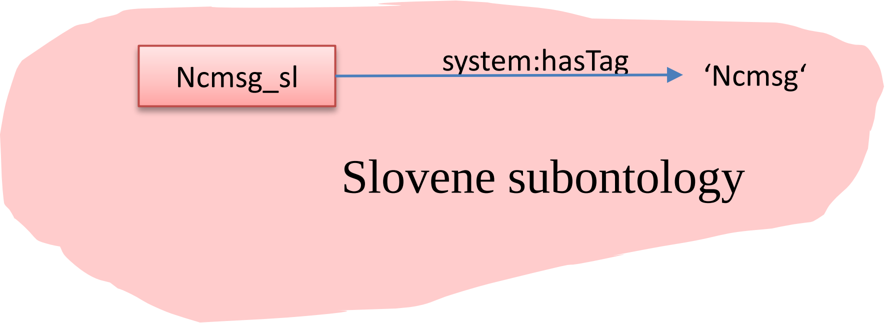


Building the MTE ontology Common specifications

8. Add examples
   - Every concept augmented examples from the language-specific specifications
9. Add definitions (manually)

Building the MTE ontology
Representing Tags

- For every language, the tags are represented in a separate language-specific subontology
  - Import common specifications
- Individuals represent tags, e.g., `Ncmsg_sl`
  - tag Ncmsg in Slovene tagset
  - Property `system:hasTag` assigns string value



Ncmsg_sl

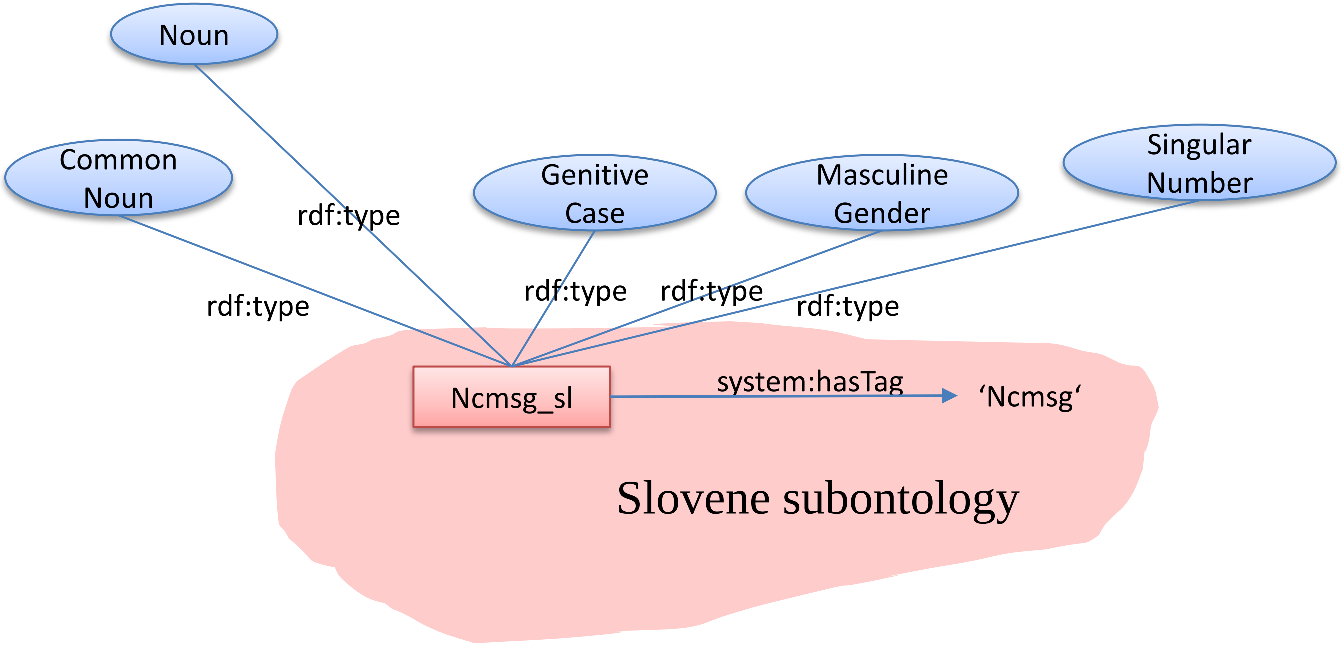
system:hasTag

‘Ncmsg‘

Slovene subontology

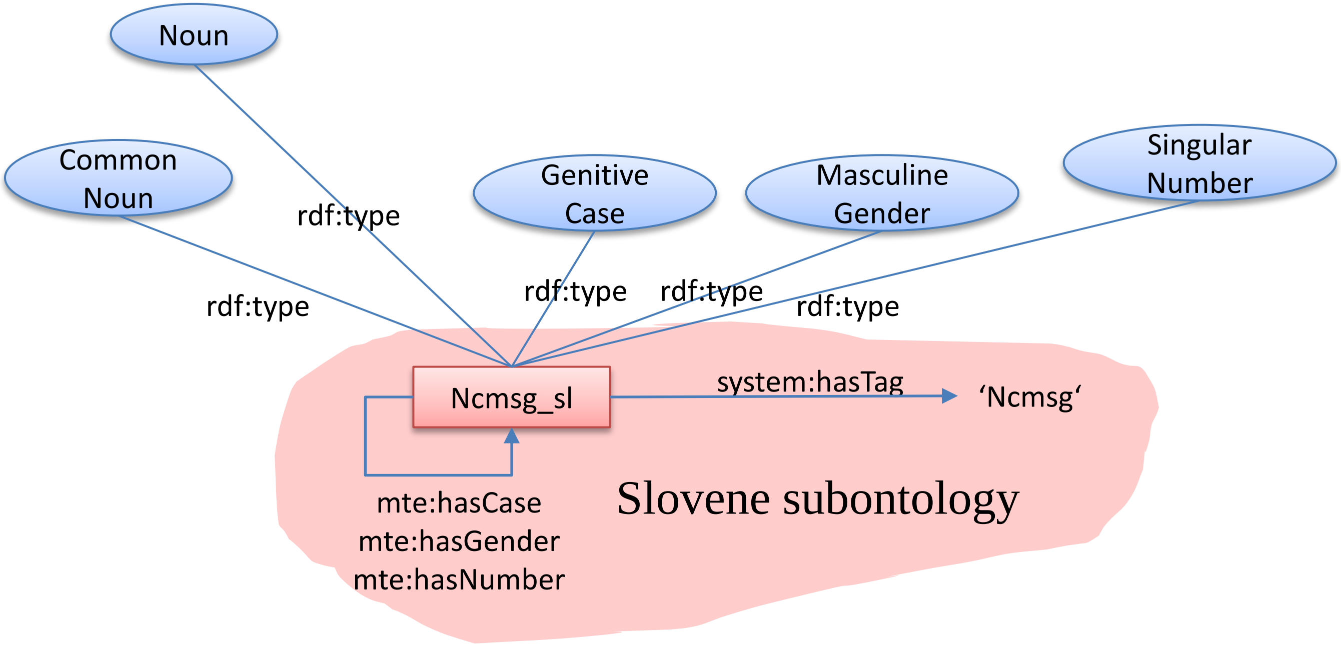
Building the MTE ontology
Representing Tags

- Individuals represent tags, e.g., `Ncmsg_sl`
  - Instance of all `MorphosyntacticCategory`s and `MorphosyntacticFeature`s expressed by the tag

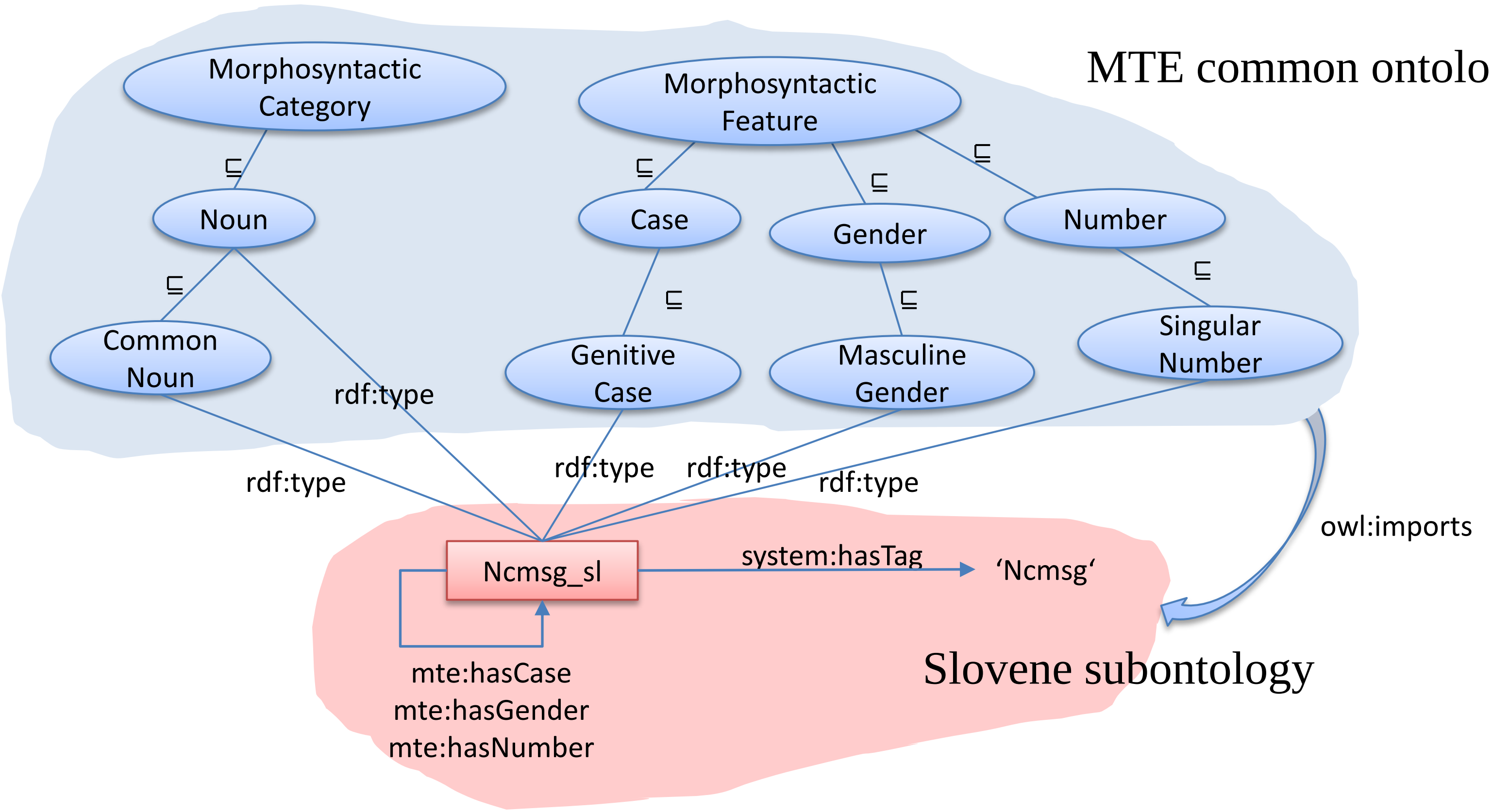


Building the MTE ontology Representing Tags

- Individuals represent tags, e.g., `Ncmsg_sl`
  - For every `MorphosyntacticFeature`, the individual is assigned the corresponding property with itself as object



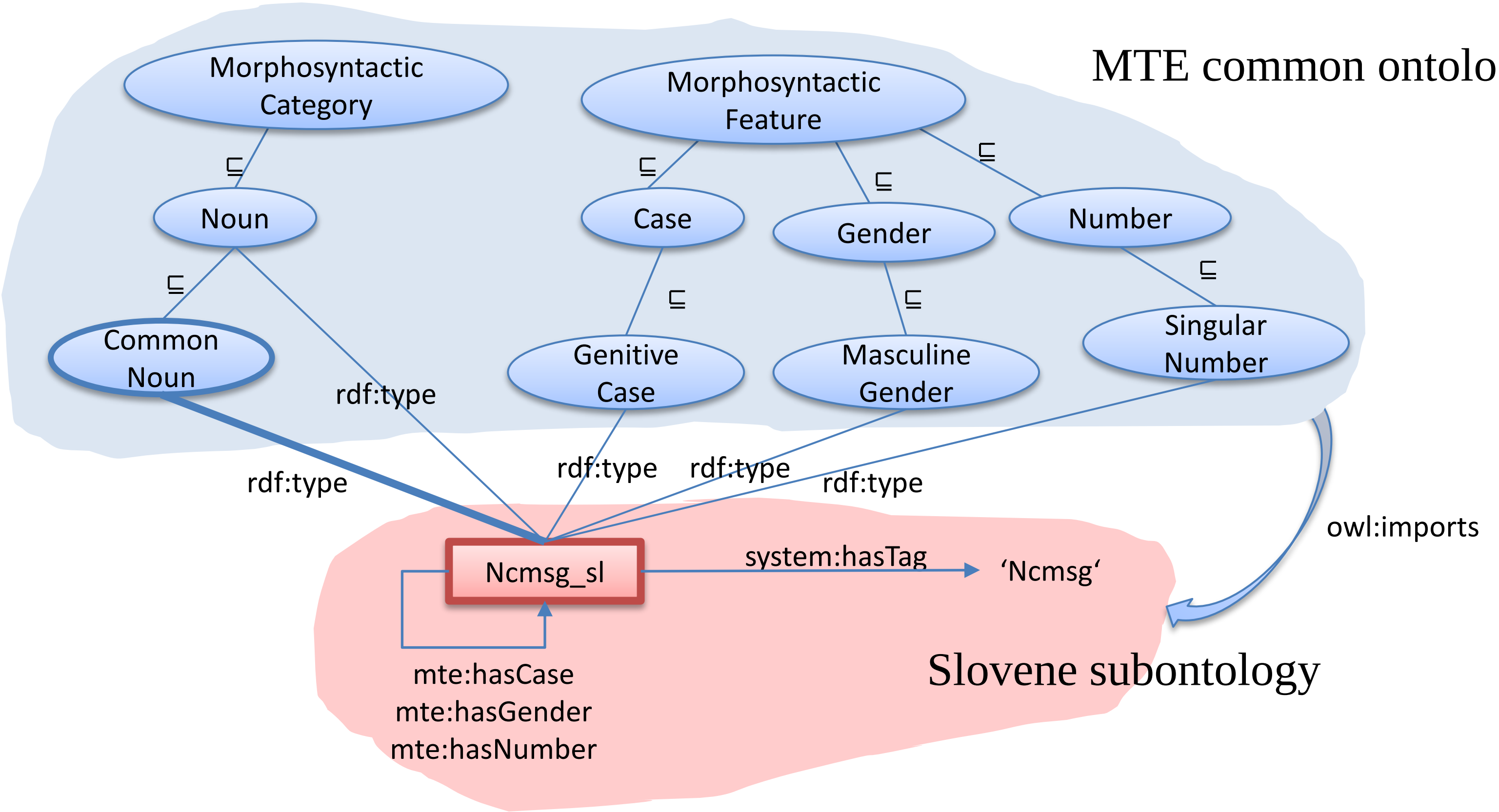
Building the MTE ontology
Representing Tags



⊑ owl:subClassOf

Building the MTE ontology
Querying Tags with OWL/DL

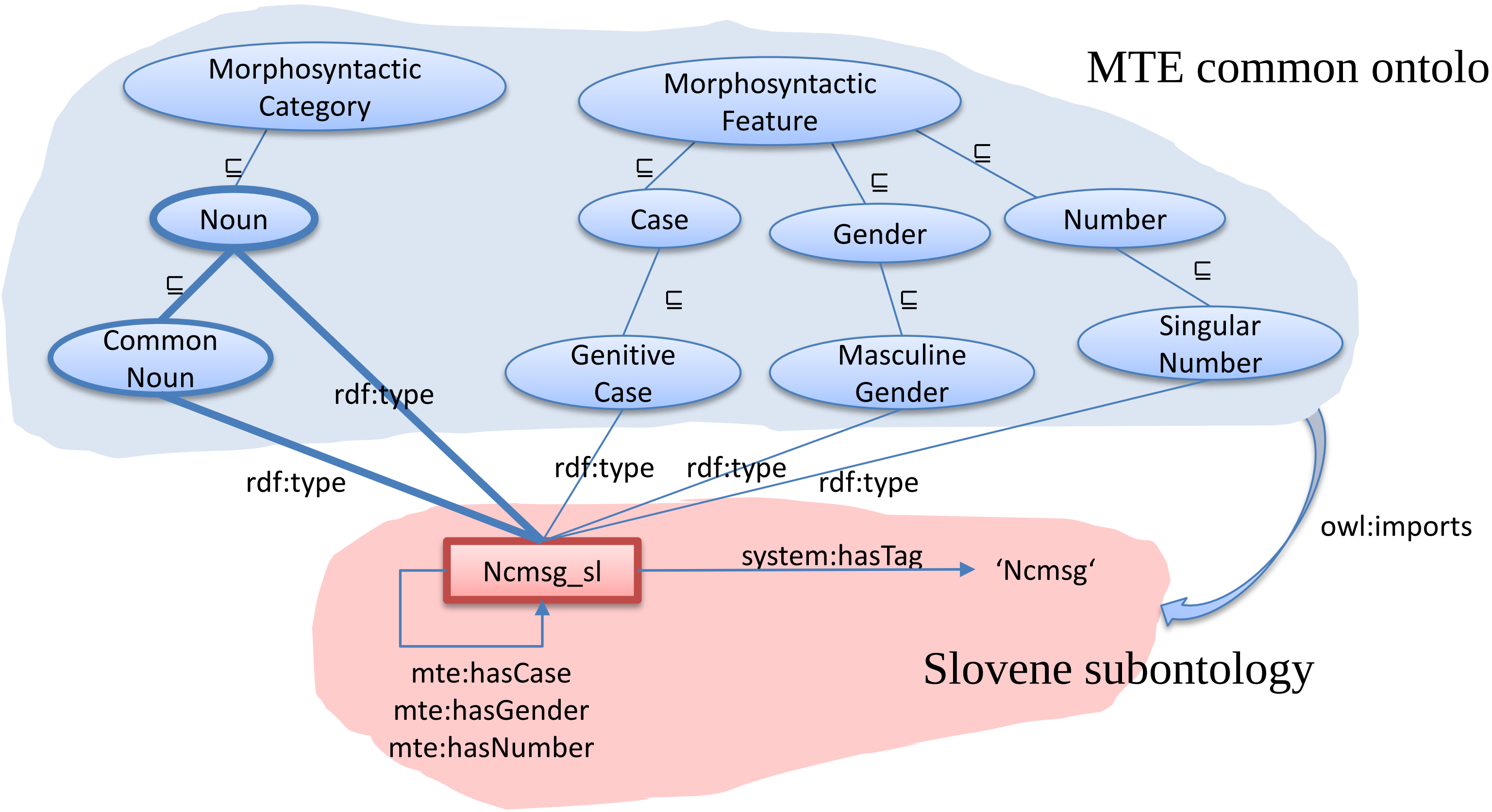
CommonNoun



⊑ owl:subClassOf

Building the MTE ontology
Querying Tags with OWL/DL

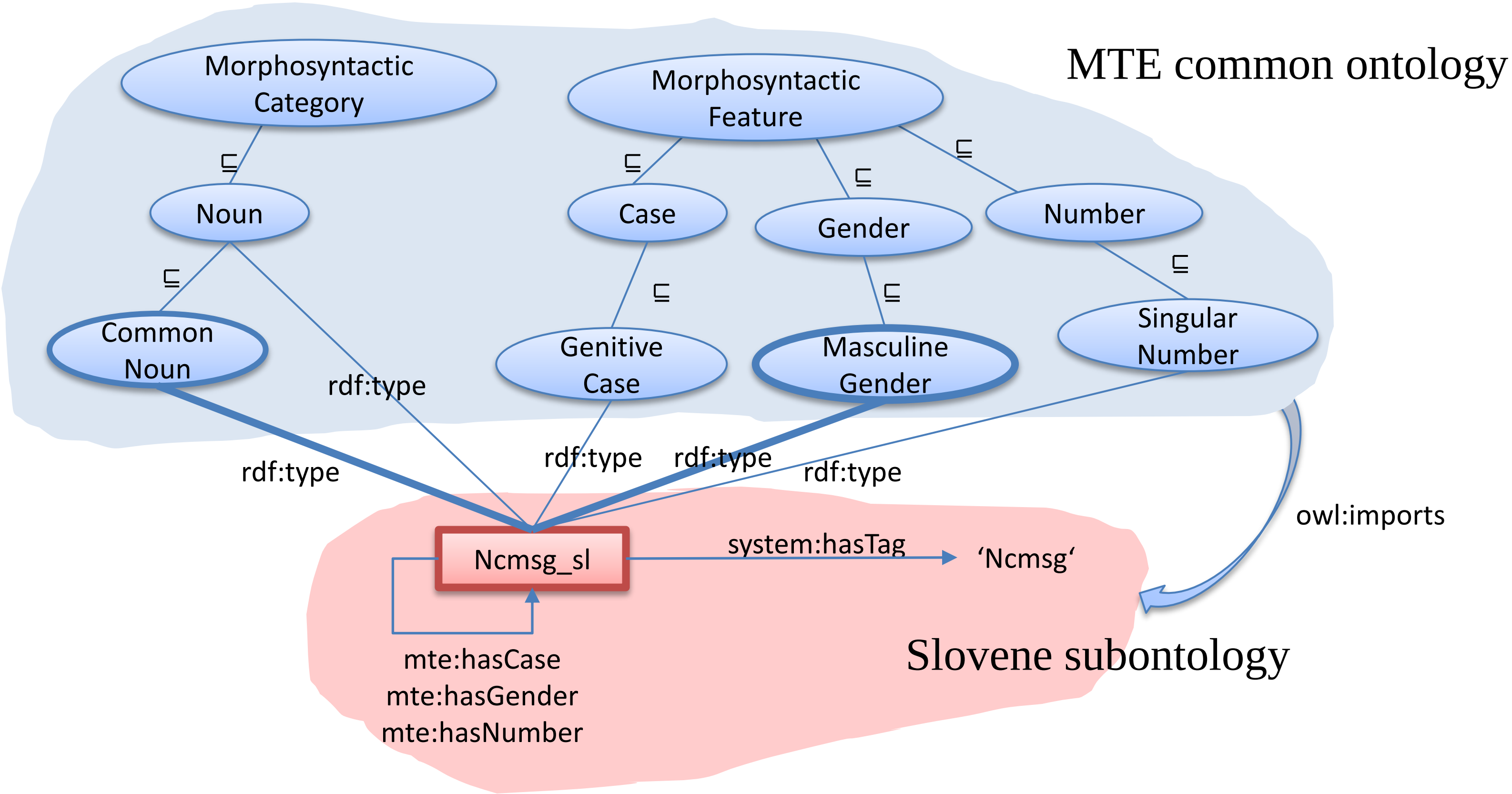
Noun



⊑ owl:subClassOf

Building the MTE ontology Querying Tags with OWL/DL

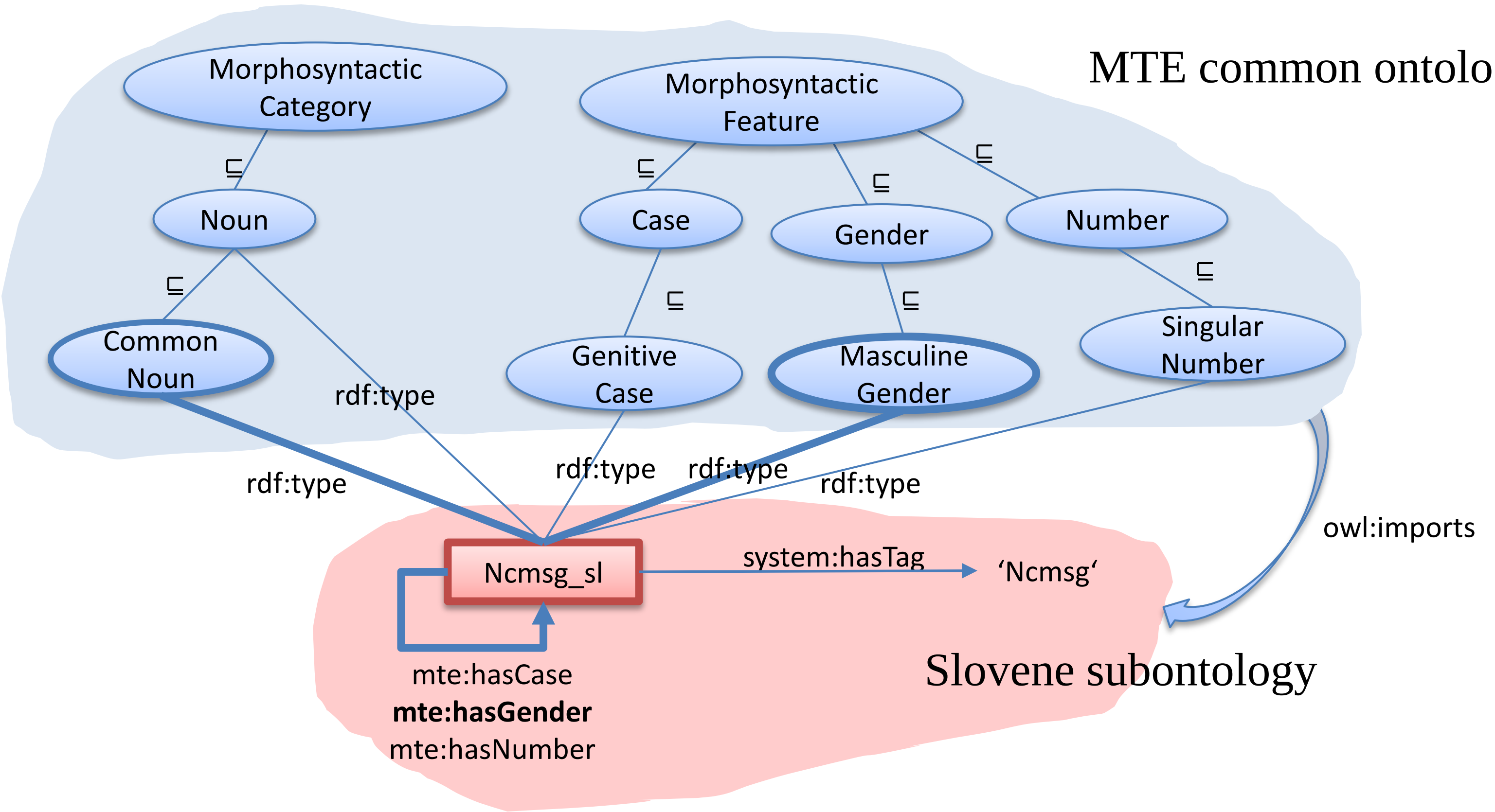
`CommonNoun and MasculineGender`



⊑ owl:subClassOf

Building the MTE ontology
Querying Tags with OWL/DL

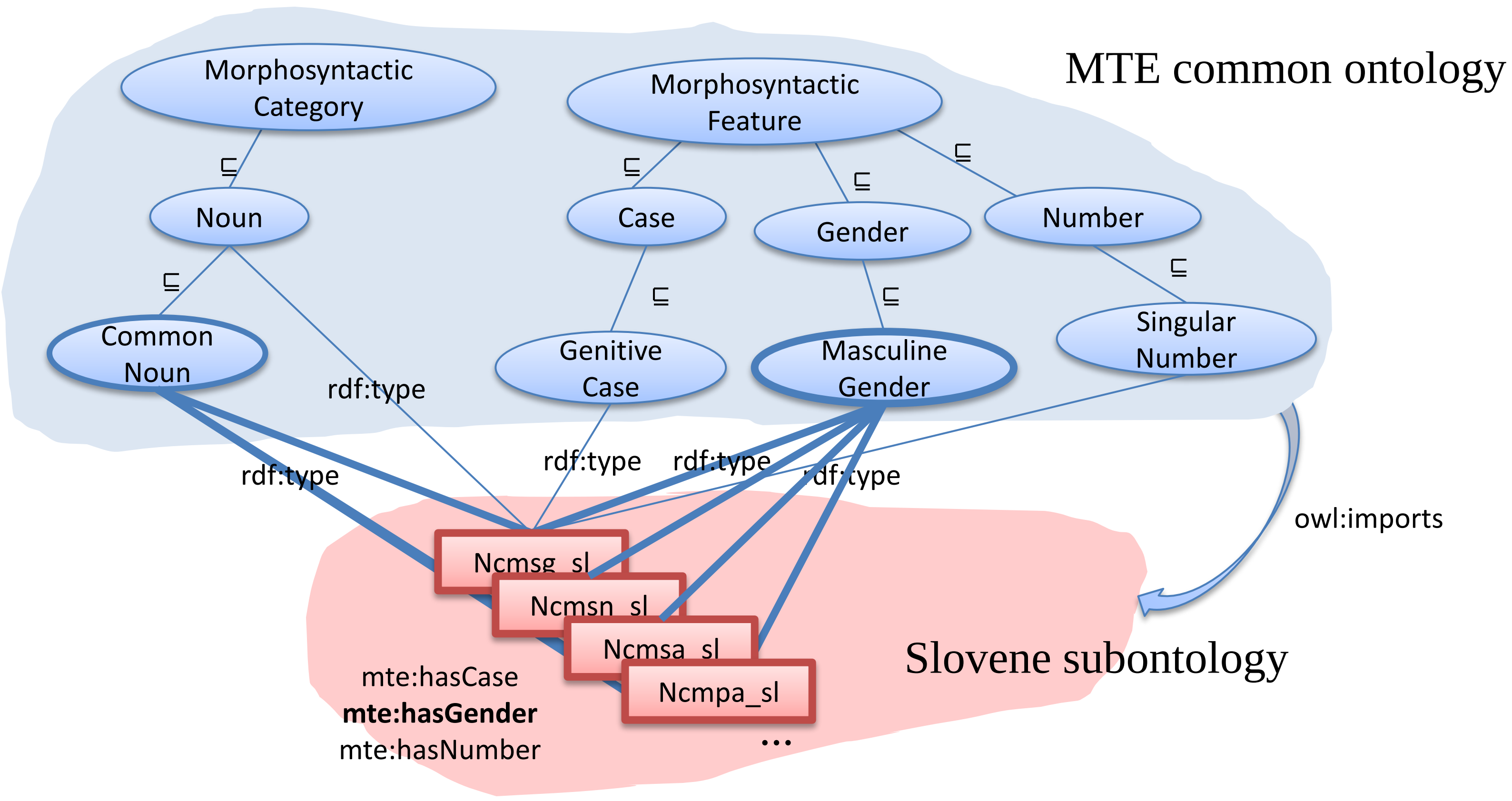
```
CommonNoun and hasGender some MasculineGender
```



⊑ owl:subClassOf

Building the MTE ontology
Querying Tags with OWL/DL

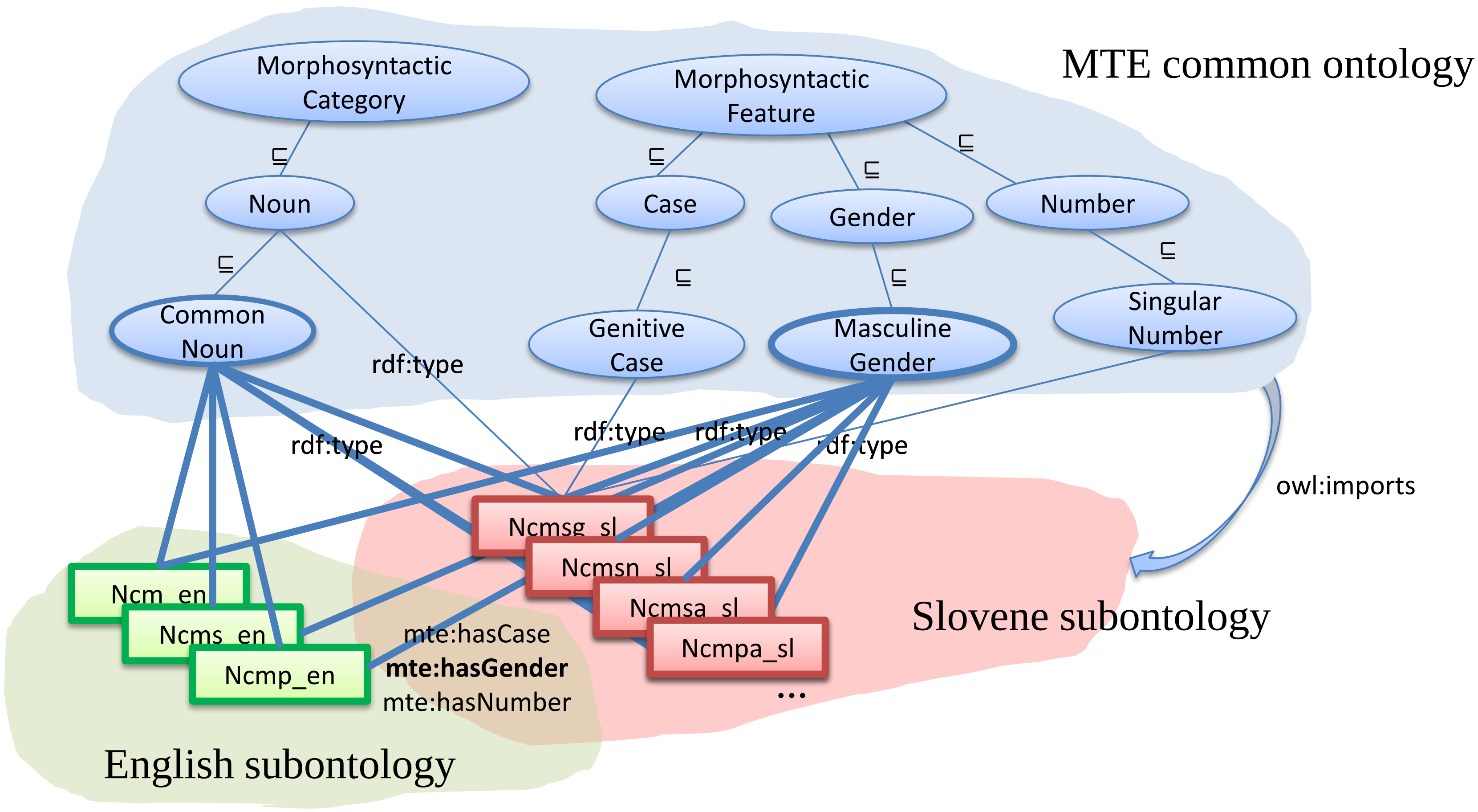
`CommonNoun and hasGender some MasculineGender`



⊑ owl:subClassOf

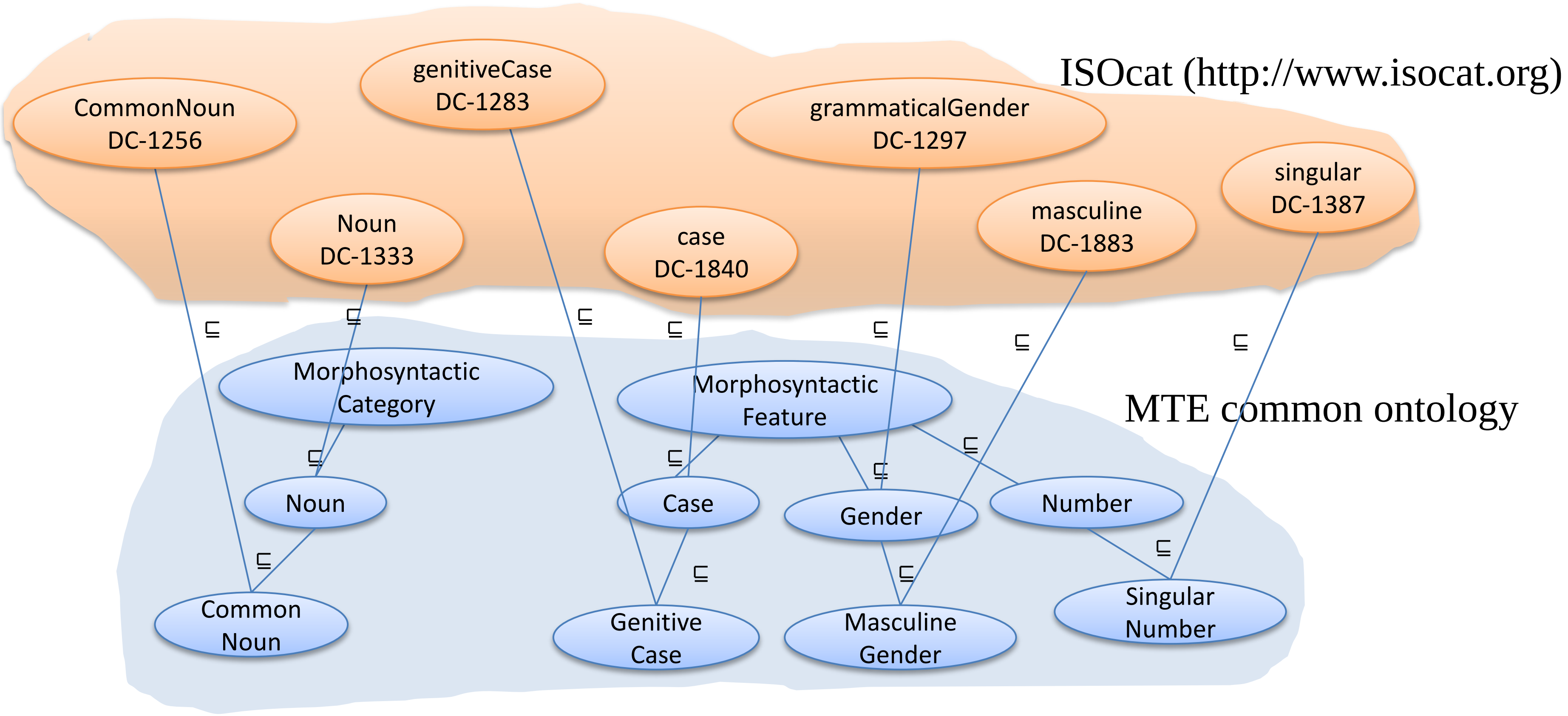
Building the MTE ontology
Querying Tags with OWL/DL

`CommonNoun and hasGender some MasculineGender`



⊑ owl:subClassOf

Towards Interoperability
Linking with terminology repositories*

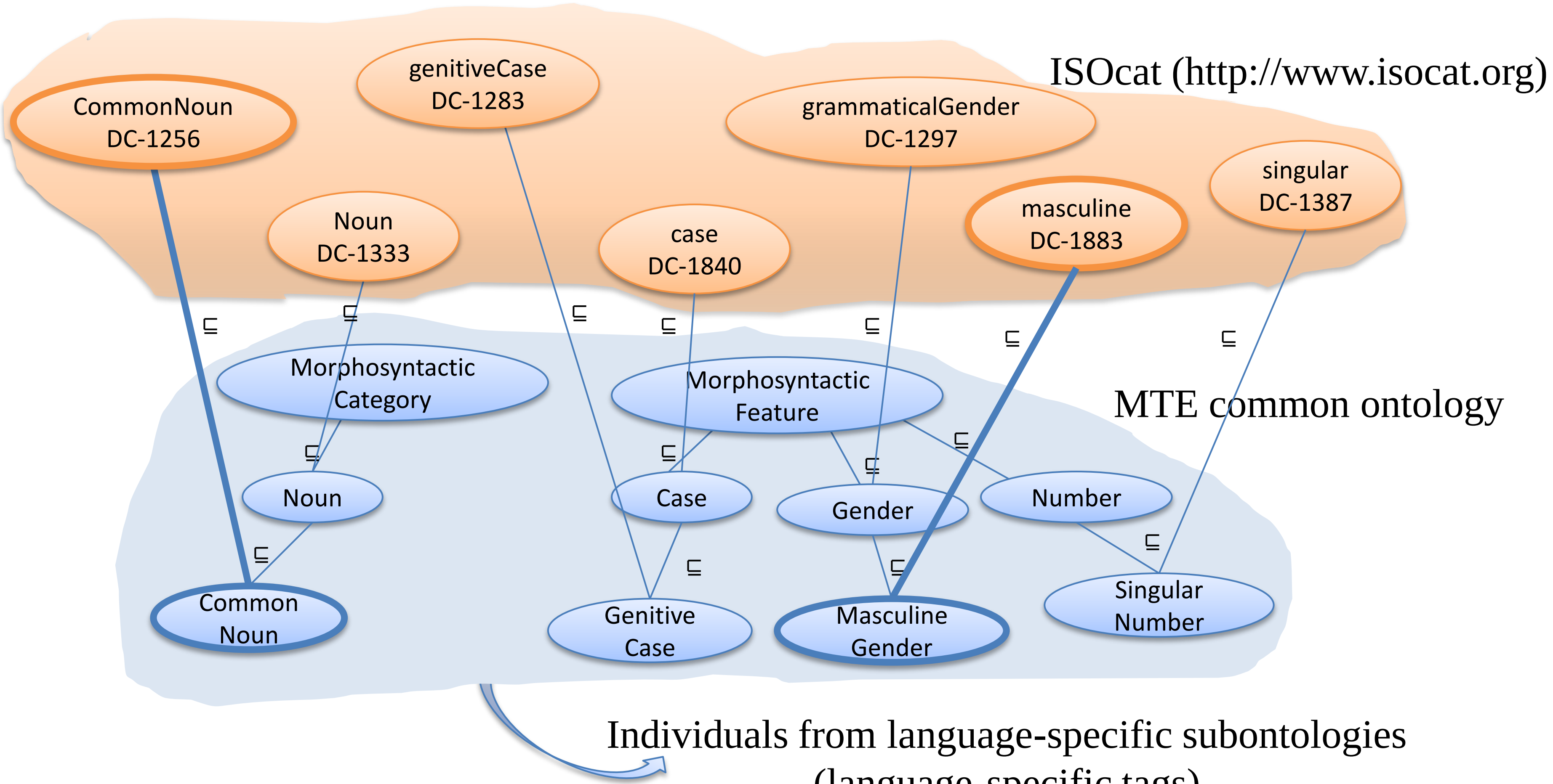


* For the approach cf. Chiarcos (2010), for the MTE, this still remains to be done

⊑ owl:subClassOf

Towards Interoperability
Querying with reference categories

`dcr:CommonNoun and dcr:masculine`

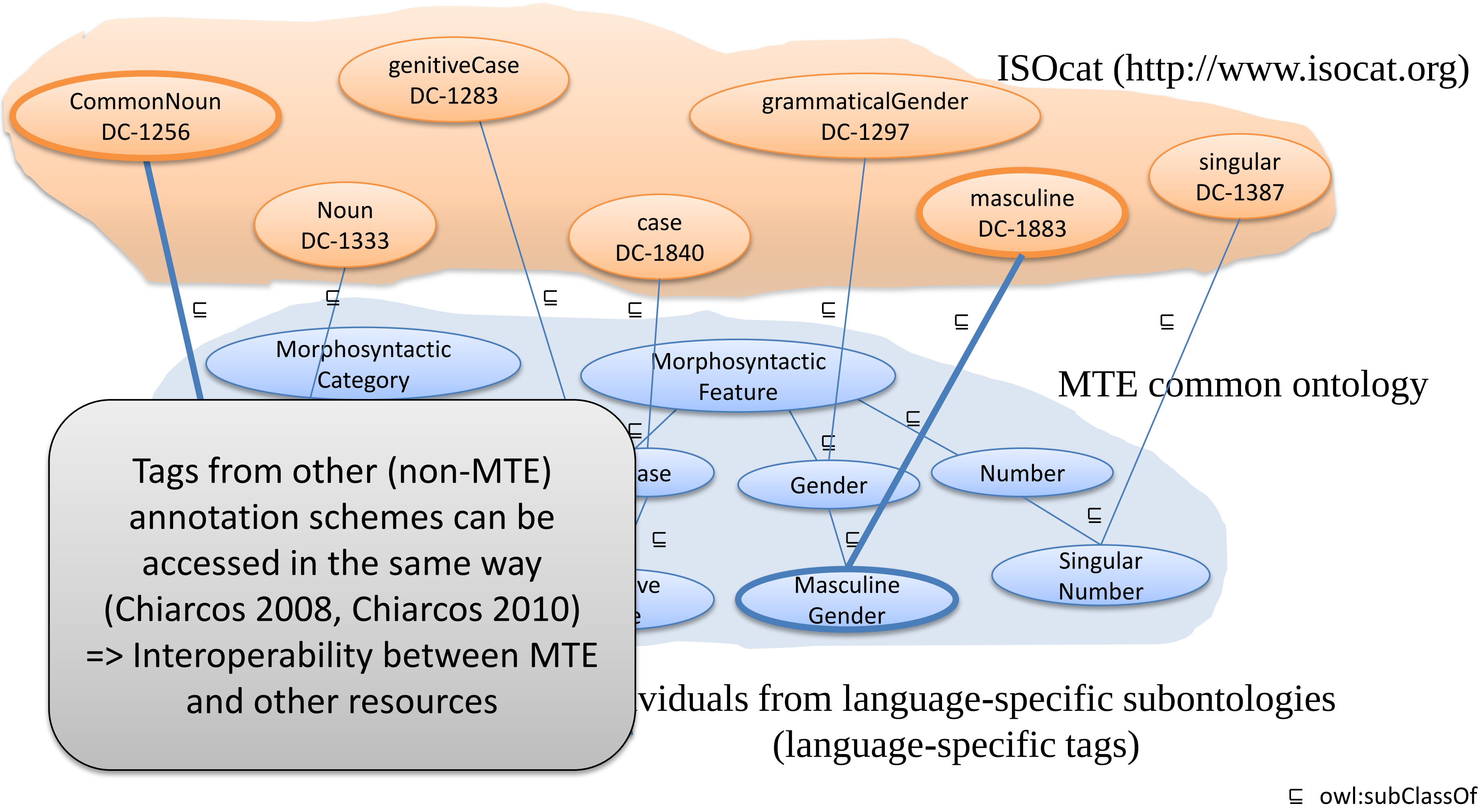


Individuals from language-specific subontologies
(language-specific tags)

⊑ owl:subClassOf

Towards Interoperability
Querying with reference categories

`dcr:CommonNoun and dcr:masculine`



Tags from other (non-MTE) annotation schemes can be accessed in the same way (Chiarcos 2008, Chiarcos 2010)
=> Interoperability between MTE and other resources

Towards Interoperability
Linking with terminology repositories

- Documentation
  - Formal and comparable specification of annotation schemes
- Cross-resource corpus querying and evaluation

  Chiarcos et al. (2008), Rehm et al. (2008)
- Combining tools with different annotation schemes (NLP pipelines, ensemble combination)

  Buyko et al. (2008), Chiarcos (2010)
- Representing NLP analyses for Semantic Web applications

  Aguado de Cea et al. (2004), Hellmann (2010)

Revising the MTE ontology

- Initial ontology built automatically
  - Conversion: XSLT
  - Validation: http://owl.cs.manchester.ac.uk/validator
- Trivial revisions
- Conceptual revisions

Revising the MTE ontology
Trivial revisions

- Expand abbreviations
  - `mte:CorrelatCoordConjunction`

    (Conjunction/Type=coord/Coord_Type=correlat)

    => `CorrelativeCoordinatingConjunction`

Revising the MTE ontology
Trivial revisions

- Expand abbreviations
- Simplifying concept names
  - `mte:DefinitenessYes`

    (Definiteness=yes)

    => `mte:Definite`

Revising the MTE ontology
Trivial revisions

- Expand abbreviations
- Simplifying concept names
- Structure induction
  - `mte:CliticProximalDeterminer` besides `mte:CliticDeterminer`

    => `mte:CliticProximalDeterminer` `owl:subClassOf` `mte:CliticDeterminer`

Revising the MTE ontology
Conceptual revisions

- Ontology requires definitions
  - Conversion done by non-expert for the languages under discussion
    - For deviations from EAGLES
      - E.g., Verb/Definiteness=1s2s, Numeral/Class=definite234
  - MTE publications, discussion with experts

=> A number of inconsistencies and redundancies identified

Revising the MTE ontology Inconsistencies: Attribute overload

- Problem

  The same attribute is used to express different functions

- Reason

  (a) Terms are interpreted differently

  (b) MTE is a *positional tagset*

  Long tags are uneconomic

  => Different language-specific phenomena are represented at the same position

  => Conflate two phenomena under a single attribute

Revising the MTE ontology Inconsistencies: Attribute overload

- Example

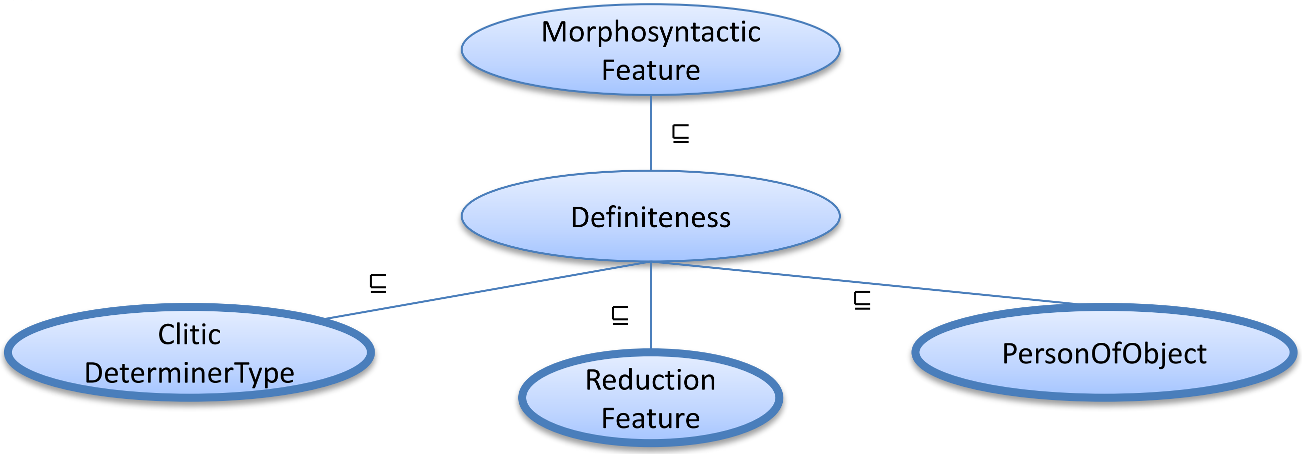
  MTE „Definiteness“

  - Type of clitic determiner (definite/indefinite/other)
    - Romanian, Bulgarian, Macedonian and Persian nouns and adjectives
  - Full or reduced adjective inflection
    - Most Slavic languages (e.g., красн**ое** vs. красн**о** in Russian)
  - Agreement with the direct object
    - „definite conjugation“ of Hungarian verbs
- Solution
  - Introduce subconcepts of Definiteness

Revising the MTE ontology Inconsistencies: Attribute overload

- Example

  MTE „Definiteness“



- Solution
  - Introduce subconcepts of Definiteness

⊑ owl:subClassOf

Revising the MTE ontology
Inconsistencies: Value overload

- Problem

  The same attribute value is used to express different functions

- Reason

  (a) Terms are interpreted differently

  (b) Avoid introducing novel features when adding a new language to MTE

Revising the MTE ontology
Inconsistencies: Value overload

- Example

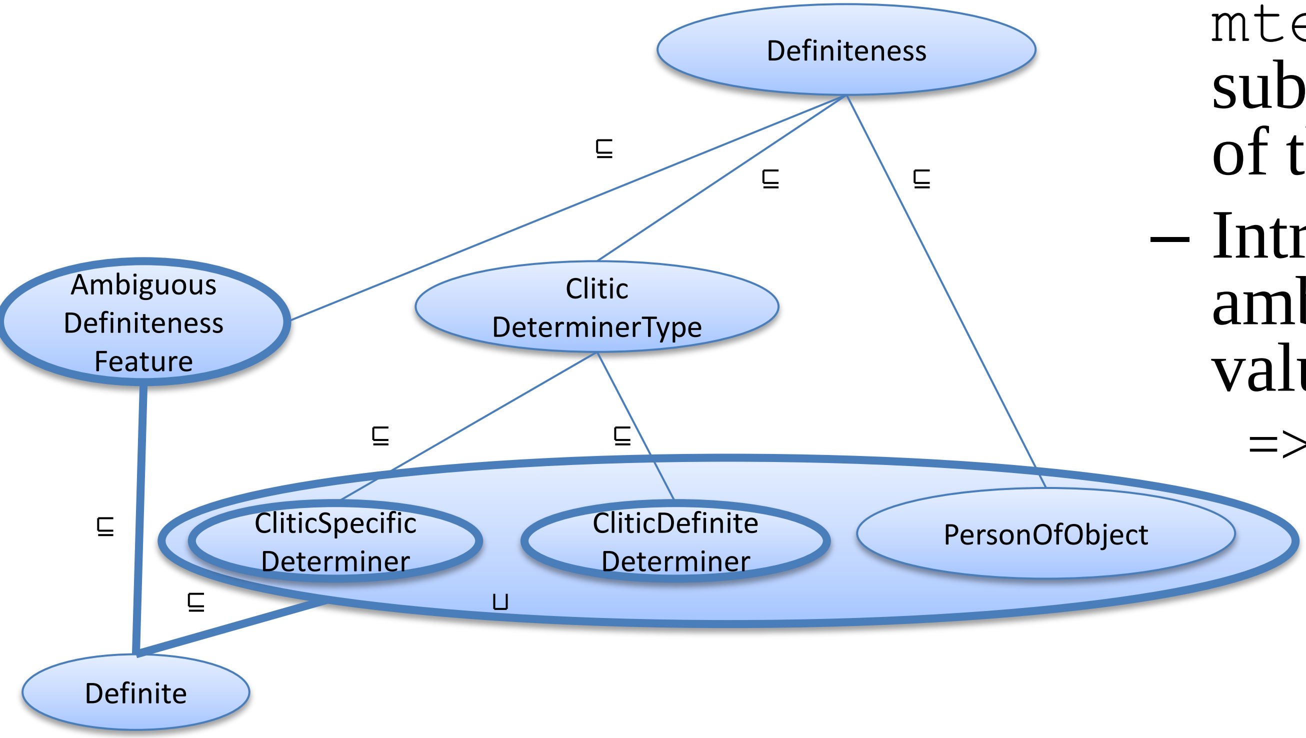
  MTE Definiteness=yes

  - Presence of a clitic definite determiner
    - Romanian, Bulgarian, Macedonian noun and adjective
  - Presence of a clitic determiner that expresses specificity
    - Persian noun and adjective
  - Verb followed by definite object argument
    - Hungarian

Revising the MTE ontology
Inconsistencies: Value overload

- Example

  MTE Definiteness=yes


Definiteness
⊑
⊑
⊑
Ambiguous Definiteness Feature
Clitic DeterminerType
⊑
⊑
⊑
CliticSpecific Determiner
CliticDefinite Determiner
PersonOfObject
⊑
⊔
Definite

- Solution
  - Establish concepts for all different senses
  - Define `mte:Definite` as a subconcept of the **join** of these novel concepts
  - Introduce concept for ambiguous feature values

    => anchor the ambiguous concept in the taxonomy

⊔ owl:join
⊑ owl:subClassOf

Revising the MTE ontology
Inconsistencies: Redundancy

- Problem

  Different attributes/values represent the same phenomenon

- Reason

  (a) Different terminological traditions

  (b) Local resolution of attribute/value overload previously existing schemes

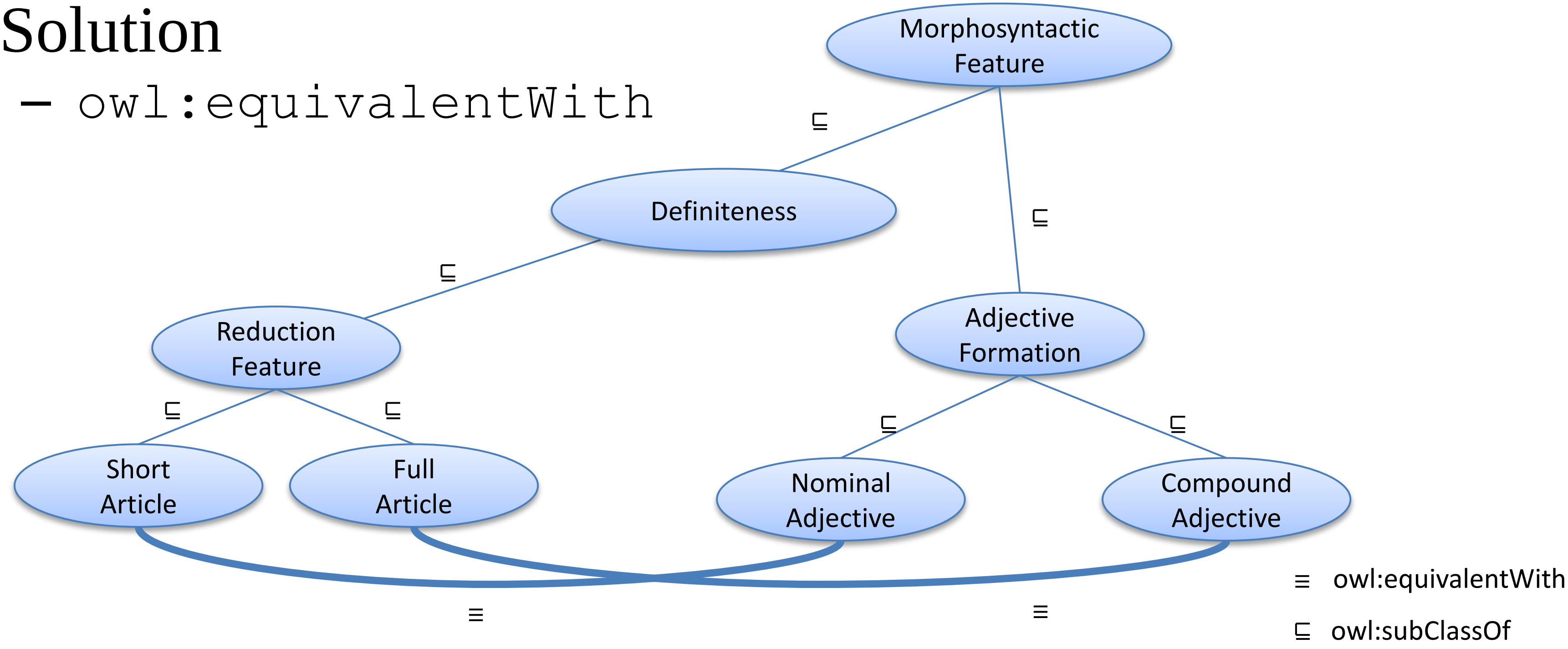
  (c) Introduction of attribute/value overload for tag set economy

Revising the MTE ontology Inconsistencies: Redundancy

- Example
  - E.g. „full“ (красн**ое**) and „reduced“ (красн**ое**) adjective inflection in Slavic languages

    Polish MTE: attribute Definiteness

    Czech MTE: attribute Formation
- Solution
  - `owl:equivalentWith`



Achievements

- OWL/DL ontologies for morphosyntactic categories for 16 languages
  - http://nl.ijs.si/ME/owl (CC BY 3.0)
  - – Top-down perspective on the MTE specs
- Conceptual problems identified and documented
  - – Documented together with morphosyntactic specifications, partially resolved
- Classification of conceptual problems and resolution strategies

Perspectives

- Can be used to link MTE with ISOcat and GOLD
  - Interoperability of MTE resources
  - Extension/revision of ISOcat/GOLD
- Can be used to guide the revision of MTE v4
  - Resolving overload and redundancy
- Process could be applied to other resources with similar benefits to be expected
  - Not that expensive to build
    - 4 days modeling and conversion; plus discussions