Distant Reading

COST
EUROPEAN COOPERATION
IN SCIENCE AND TECHNOLOGY

Annotating Named Entities in TEI

Virtual Training School on NER

COST Action CA 16204 « Distant reading »

23rd of March 2021

PRESENTATION PLAN

- What are the TEI tags for named entities?
- How to use the tags – from a simple to a more elaborated annotation
- Converting txt files into TEI/XML files

TEI about Names, Dates, People, Places

- 2 sections:

- Section 3.5 « Names, Numbers, Dates, Abbreviations, and Addresses » within chapter 3 « Elements available in all TEI documents »

- Chapter 13: more in-depth encoding of « Names, Dates, People and Places »

- Most of the above are « referring strings », that can be annotated with « referencing strings », aka <rs>

TEI simple annotation

<rs type=‘PERS’>Harry Potter</rs>

<rs type=‘ROLE’>Her Majesty</rs>

<rs type=‘ROLEMISC’>his cousin</rs>

<rs type=‘ORG’>the Ministry of Magic</rs>

<rs type=‘PLACE’>Hogwarth</rs>

<rs type=‘WORK’>The Tales of Beedle the bard</rs>

<rs type=‘EVENT’>the Yule Ball</rs>

<rs type=‘FAC’>Astronomy tower</rs>

<rs type=‘MISC’>Monday</rs>

<rs type=‘DEMO’>the Griffindors</rs>

TEI simple annotation

- Do we keep the values of @type identical to those used in BRAT? (typography, expressivity, alignment with other taxonomies)
- <rs> is not the only tag recommended by TEI
- <rs> can nest, but in some cases (question/response split over two paragraphs) the nesting may be problematic
- How do we link entities pointing to the same referent?

TEI annotation: types of NE

- Competing classifications (see NER campaigns)
- Values of @type are not constrained in TEI > you can define your own when customizing the schema
- NER on literary texts: problem of the purely fictitious/ mixed characters, places, events, roles, etc.

1. <rs type=‘PERS’>Harry</rs> stopped under the big statue of <rs type=‘PERS’>Queen Victoria</rs>.
2. <rs type=‘litchar’>Harry</rs> stopped under the big statue of <rs type=‘histchar’>Queen Victoria</rs>.
3. <rs type=‘PERS’>Harry</rs> stopped under <rs type=‘WORK’>the big statue of <rs type=‘ROLE’>Queen</rs> <rs type=‘PERS’>Victoria</rs></rs>.

TEI annotation: types of NE

- Whatever the choice, the values of @type should be documented in the header

<tagUsage gi=‘rs’>This element is used for marking up all named entities in the text. The type attribute on on this element takes one or more of the following values:

<list type="gloss">

<label>PERS</label>

<item>Proper names of people including first names, last names, individual or family names, fictional names and unique nicknames. This applies also to gods. Generational markers (Jr., VIII), and royal titles (Queen, Sir are included). Honorific titles (Mr., Mrs., Miss, Ms, Dr, Prof) are included when they occur followed by the proper name. Category to be used for named animals too. </item>

<label>ORG</label>

<item>Names of companies, political parties, educational institutions, sport teams, hospitals, museums, libraries etc. Especially, hotels, museums, hospitals, libraries, churches and temples, commercial facilities, ... </item>

Etc.

</list>

TEI alternatives

Eltec BRAT annotation values	TEI potential « translations »
PERS (person names)	<name>, <persName> or <rs type=‘person’>
ROLE (profession, nobility, office)	<roleName> or <rs type=‘role’>
ROLEMISC (family, epithet)	<rs type=‘family’>for the first, <roleName> for the second
ORG (organizations, political parties, companies…)	<orgName>, <name type=‘organisation’> or <rs type=‘organisation’>
PLACE (all types of locations)	<placeName>, <geogName>, <name type=‘place’> or <rs type=‘place’>
WORK (titles of art works)	<title> or <rs type=‘title’>
EVENT	<rs type=‘event’> (not <event>!!!)
FAC (infrastructure, superstructure, transportation)	<rs type=‘facility’>
MISC (dates, times, measures)	<date>, <time>
DEMO (names of kinds of people)	<rs type=‘demonym’>

TEI alternatives: person names

- Harry Potter

```
<persName>
        <forename>Harry</surname>
        <surname>Potter</surname>
</persName>
```

- Tom Riddle, Lord Voldemort

```
<persName>
        <forename>Tom</forename>
        <surname>Riddle</surname>
        <addName>lord Voldemort</addName>
</persName>
```

Or

```
<persName>
        <forename>Tom</forename>
        <surname>Riddle</surname>
        <roleName>lord</addName>
        <addName>Voldemort</addName>
</persName>
```

TEI alternatives: roles

- Barnabas the Barmy

BRAT: <PERS>Barnabas</PERS> <ROLEMISC>the Barmy</ROLEMISC>

TEI: <persName> <forename>Barnabas</forename> <roleName>the Barmy</roleName> </persName>

- his mother Lily

BRAT: <ROLEMISC>his mother</ROLEMISC> <PERS>Lily</PERS>

TEI: <rs type=‘family’>his mother</rs> <name>Lily</name>

Or

<rs type=‘family’>his mother</rs> <persName><forename>Lily</foreName></persName>

TEI alternatives: place names

- Little Whinging, Surrey

```
<placeName>
        <settlement>Little Whinging</settlement>
        <region>Surrey</region>
</placeName>
```

- 12, Grimauld Place, Islington, London

```
<address>
        <addrLine>12, Grimauld Place</addrLine>
        <addrLine>Islington</addrLine>
        <addrLine>London</addrLine>
</address>
```

TEI alternatives: place names

- A playground 200 m far from Little Whinging

```
<rs>a playground</rs> <placeName>
                        <measure>200 m</measure>
                        <offset>far from</offset>
                        <settlement>Little Whinging</settlement>
                    </placeName>
```

- the Portkey placed at the top of Stoatshead Hill

```
The Portkey placed <placeName>
                        <offsett>at the top of</offset>
                        <geogName>
                        <name>Stoatshead</name> <geogFeat>Hill</geogFeat>
                        </geogName>
                   </placeName>
```

TEI alternatives: events

- He wanted to invite Cho Chang to the Yule ball

He wanted to invite <name>Cho Chang</name> to <rs type=‘event’>the Yule ball</rs>

⚠️ Not <event> (alas!)

> « An <event> element is usually used to record information about a place, or a person; for this reason the element usually appears as content of a <place> or <person>. However, it is also possible to describe events independently of either a person or a place. This may be useful in such applications as chronologies, lists of significant events such as battles, legislation, etc. » (TEI Guidelines, 13.3.4.3)

TEI alternatives: dates, time, measures

- BRAT: value MISC for all

TEI

<date>: « contains a date in any format »

- unsure however about:

<date>That Monday</date> he waked up full of determination.

<time>: « contains a phrase defining a time of day in any format »

<time when="08:48:00">8:48</time>

<date when="2001-09-11T12:48:00">Sept 11th, 12 minutes before 9 am</date>

<time when="1999-01-04T20:42:00-05:00">4 janvier 1999 à 8h de l'après-midi.</time>

<measure>

<measure><num>2</num> <unit>galleons</unit></measure>

TEI alternatives: works of art

<title>: works fine for texts, sometimes music and paintings, but it seems more of a tag abuse for statues and other works

Last year, we read <title>Le Rouge et le Noir</title>.

*He was seated under <title>Molière’s bust</title>.

He was seated under <rs type=‘work’>Molière’s bust</rs>

Enriching the TEI annotation

- Referencing strings pointing to the same entity

  @key

  <rs type='litchar' key='HP'>Harry</rs> stopped. Snape was watching him. <rs type='litchar' key='HP'>Potter</rs>, he said, …

  - Problem is to remain consistent about the @key values, since they are not linked to an xml:id
  - One can also use a standard reference, documented using a <taxonomy> element in the TEI header

- Referencing strings pointing to the same entity, defined elsewhere on the web

  @ref

  <rs type='litchar' ref='https://en.wikipedia.org/wiki/Harry_Potter_(character)'>Harry</rs> stopped. <rs type='litchar' ref='https://en.wikipedia.org/wiki/Severus_Snape'>Snape</rs> was watching him.

Enriching the TEI annotation

- How named entities are named?
- Single NE/ group NE

Enriching the TEI annotation

ISO/FDIS 24617-9 **Language resource management — Semantic annotation framework —** Part 9: **Reference annotation framework (RAF)**

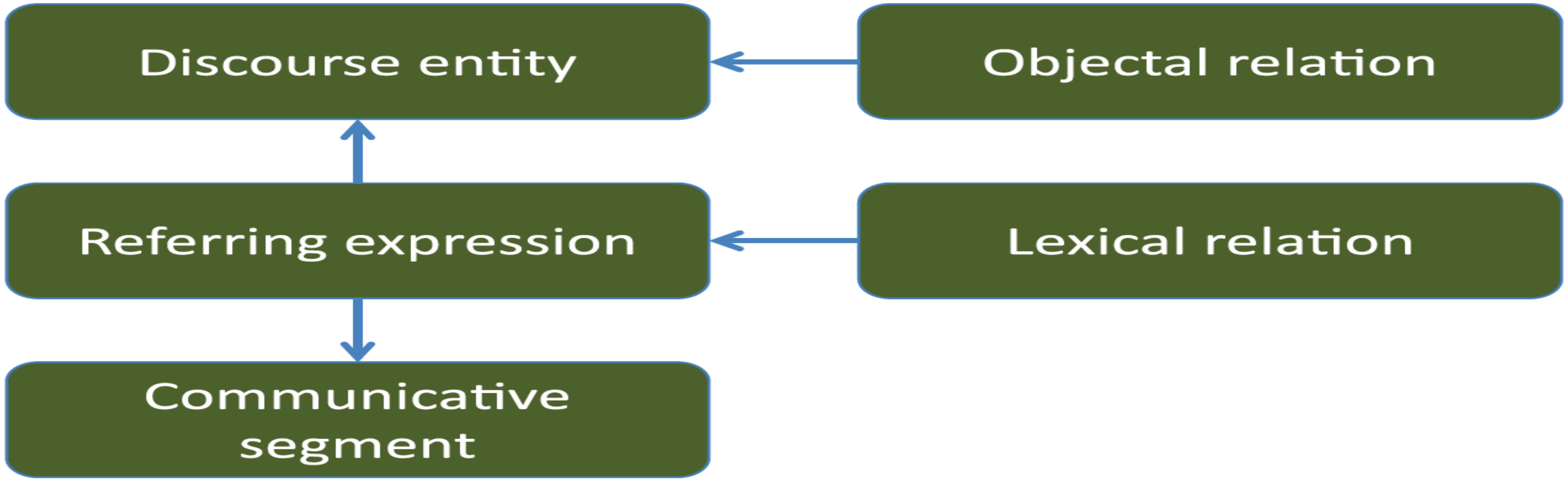


Figure 1 — Meta model for reference annotation

Enriching the TEI annotation

1° Split text in tokens

```
<w xml:id="w1">The</w>
<w xml:id="w2">City</w>
<w xml:id="w3">is</w>
<w xml:id="w4">a</w>
<w xml:id="w5">tiny</w>
<w xml:id="w6">part</w>
<w xml:id="w7">of</w>
<w xml:id="w8">the</w>
<w xml:id="w9">metropolis</w>
<w xml:id="w10">of</w>
<w xml:id="w11">London</w>.
```

2° Insert <standOff> element in header.

Enriching the TEI annotation

3° Annotate referring expressions as <span> in the <standOff> element.

```
<spanGrp type="referringExpression">
    <span xml:id="m1" from="#w1" to="#w2" ana="#fs1"><!– The City --></span>
    <span xml:id="m2" from="#w4" to="#w11"  ana="#fs2"><!– a tiny part of the metropolis of London--></span>
</spanGrp>
```

- Same for London

Enriching the TEI annotation

4° Add linguistic information about the referring expression

```
<fs xml:id="fs1"> <!--The City-->
  <f name="syntacticCategory"><symbol value="nounPhrase"/></f>
  <f name="determinerType"><symbol value="definite"/></f>
  <f name="referentialStatus"><symbol value="discourseNew"/></f>
</fs>

<fs xml:id="fs2"> <!—a tiny part of the London metropolis-->
  <f name="syntacticCategory"><symbol value="nounPhrase"/></f>
  <f name="referentialStatus"><symbol value="discourseOld"/></f>
</fs>
```

(possibility to add other information: compound, contains one or several other NE, etc.)

Enriching the TEI annotation

5° Annotate discourse entities

```
<listAnnotation type="discourseEntities">
    <interp xml:id="e1" inst="#m1 #m2" type="discourseEntity">
    <!-- CITY --></interp>
    <interp xml:id="e2" inst="#m3" type="discourseEntity">
    <!-- LONDON --></interp>
</listAnnotation>
```

Enriching TEI encoding

6° Annotate lexical relations

```
<listAnnotation type= "lexicalRelations">
    <link xml:id="link1" ana="#fs3" target="#m1 #m2”
    type="lexicalRelation"/>
    <fs xml:id="fs3">
        <f name="lexicalRelation"><symbol value="meronymy"/></f>
    </fs>
</listAnnotation>
```

Types of lexical relations

sameHead

Pronominal

Synonymy

Hyponymy

Hypernymy

Compatibility

Antonymy

Incompatibility

Acronymy

Meronymy

Metonymy

Enriching TEI encoding

7° Anotate referential relations

```
<listAnnotation type="objectalRelations">
    <link xml:id="link2" ana="#fs4" target="#e1 #e2” type="objectalRelation"/>
    <fs xml:id="fs4">
        <f name="objectalRelation"><symbol value="partOff"/></f>
    </fs>
</listAnnotation>
```

Types of referential relations

objectalIdentity

partOf

Subset

memberOf

referentialDisjunction

In-text alternative: a proposal

- Annotate the referring strings : <rs>
- Annotate the entities they refer to:

  a list of some form in the header, bearing unique identifiers for each entity and providing an unified designator > @key

- Annotate the nature of the referring string

  @ana (prepare a list of values, or refer to an existing one)

- Annotate the relations between the referred entities

  @corresp

In-text alternative: creating unique identifiers for entities

```
<profileDesc>
      <particDesc>
        <listPerson>
          <person xml:id="SJ">
            <name>Sissy (Cecilia) Jupe</name>
          </person>
         </listPerson>
      </particDesc>
  </profileDesc>
```

- See chapter 16 of the Guidelines, « Linking, Segmentation and Alignment »

Enriching the TEI annotation: relations between the discourse entities

<p>‘**<rs key=‘SJ’>Girl number twenty</rs>**’, said Mr. Gradgrind, squarely pointing with his square forefinger, ‘I don’t know **<rs key=‘SJ’>that girl</rs>**. Who is **<rs key=‘SJ’>that girl</rs>**?’</p>

<p>‘**<rs key=‘SJ’>Sissy Jupe</rs>**, sir,’ explained **<rs key=‘SJ’>number twenty</rs>**, blushing, standing up, and curtseying.</p>

<p>‘Sissy is not a name,’ said Mr. Gradgrind. ‘Don’t call **<rs key=‘SJ’> yourself</rs> <rs key=‘SJ’>Sissy</rs>**. Call **<rs key=‘SJ’>yourself</rs> <rs key=‘SJ’>Cecilia</rs>**.’</p>

<p>‘It’s father as calls **<rs key=‘SJ’>me</rs> <rs key=‘SJ’>Sissy</rs>**, sir,’ returned **<rs key=‘SJ’>the young girl</rs>** in a trembling voice, and with another curtsey.</p>

Enriching the TEI annotation: relations between the discourse entities

<p>‘<rs key=‘SJ’ ana=‘#hyperonym #identifier’>Girl number twenty</rs>’, said Mr. Gradgrind, squarely pointing with his square forefinger, ‘I don’t know <rs key=‘SJ’ ana=‘#hyperonym’>that girl</rs>. Who is <rs key=‘SJ’ ana=‘#hyperonym’>that girl</rs>?’</p>

<p>‘<rs key=‘SJ’ ana=‘#fullName’>Sissy Jupe</rs>, sir,’ explained <rs key=‘SJ’ ana=‘#identifier’>number twenty</rs>, blushing, standing up, and curtseying.</p>

<p>‘Sissy is not a name,’ said Mr. Gradgrind. ‘Don’t call <rs key=‘SJ’ ana=‘#pronoun’>yourself</rs> <rs key=‘SJ’ ana=‘#forename’>Sissy</rs>. Call yourself <rs key=‘SJ’ ana=‘#altname’>Cecilia</rs>.’</p>

<p>‘It’s father as calls <rs key=‘SJ’ ana=‘#pronoun’>me</rs> <rs key=‘SJ’ ana=‘#forename’>Sissy</rs>, sir,’ returned <rs key=‘SJ’ ana=‘#hyperonym’>the young girl</rs> in a trembling voice, and with another curtsey.</p>

Enriching the TEI: objectal relations

- In the teiHeader

```
<sourceDesc>
    <list>
        <item xml:id="LG">Louisa Gradgrind</item>
        <item xml:id ="TG">Thomas Gradgrind</item>
    </list>
 </sourceDesc>
```

- See chapter 16 of the Guidelines, « Linking, Segmentation and Alignment »: <title>, <name>, <seg> in the body + <linkGrp> in the back (or other document)

Enriching the TEI: objectal relations

<p>There were <rs key="Gradgrinds" type="persGroup" ana="#identifier #hyperonym #name" corresp="#LG #TG" subtype="partial">**five young Gradgrinds**</rs>, and <rs key="Gradgrinds" type="persGroup" ana="#pronoun" corresp="#LG #TG" subtype="partial">they</rs> were models every one. [...]</p>

<p>A space of stunted grass and dry rubbish being between him and the young rabble, he took his eyeglass out of his waistcoat to look for <rs key="child" type="pers" ana="#hyperonym" corresp="#LG #TG" subtype="partial">**any child he knew by name, and might order off**</rs>. Phenomenon almost incredible though distinctly seen, what did he then behold but <rs key="LG" type="pers" ana="#name #qualifier">**his own metallurgical Louisa**</rs>, peeping with all her might through a hole in a deal board, and <rs key="TG" type="pers" ana="#name #qualifier">**his own mathematical Thomas**</rs> abasing himself on the ground to catch but a hoof of the graceful equestrian Tyrolean flower-act!</p>

Transition

RECOGITO

Exercise 1: your samples (txt files)

- Download your sample
- Connect to http://brat.jerteh.rs/
- Perform NER with Spacy and/ or Stanford
- Download the result in xml
- Compare the result with the LitBank annotation (folder « novels »)

Exercise 1: your samples (txt files)

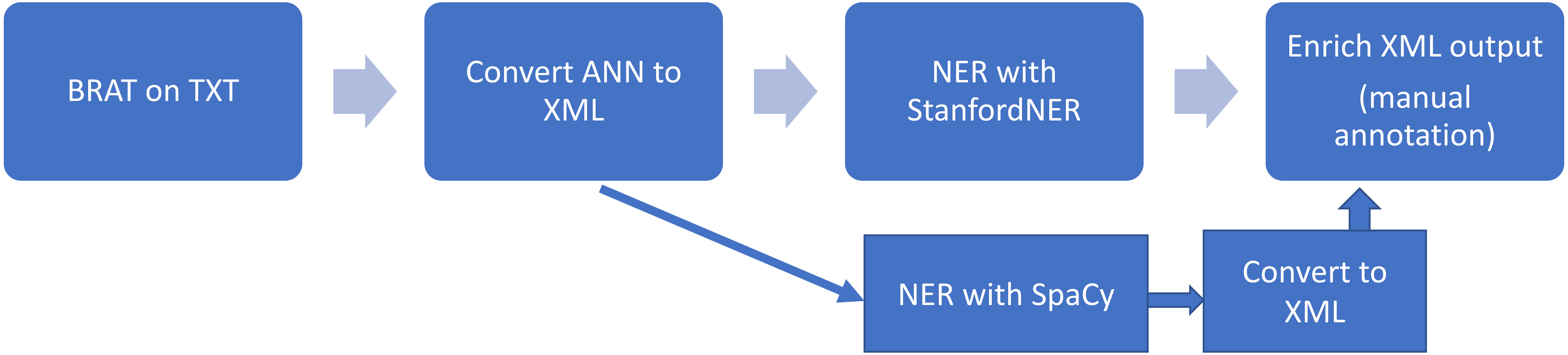
Great expectations

Natalja Tomic, Sample 1
Vojtech Malinek, Sample 2
Anders Malvik, Sample 3
Lovro Skopljanac, Sample 4
Katie Mishler, Sample 5
Katrin Horn, Sample 6
Lauren Cassidy, Sample 7
Andrejka Zejn, Sample 8
Luiza Marinescu, Sample 9
Fotini Kaidaki, Sample 10

Jane Eyre

Jana-Katarina Mende, Sample 11
Susana Sotelo Docio, Sample 12
Marek Debnar, Sample 13
Ioana Lionte, Sample 14
Alexandra Olteanu, Sample 15
Lucretia Pascariu, Sample 16
Cezary Rosinski, Sample 17
Lucija Mandic, Sample 18
Lou Burnard, Sample 19

NER workflow



BRAT on TXT

Convert ANN to XML

NER with StanfordNER

Enrich XML output

(manual annotation)

NER with SpaCy

Convert to XML

Exercise 2: manual NE annotation

- Download your sample
- Identify the named entities and annotate them with <rs>
- For each different NE, generate a @key (pay attention to use the same key for the <rs> pointing towards the same NE)
- Add a @type with one of the following values: pers, persGroup, demo, place, event, fac, work, other
- Add a @ana with one of the following values: name, fullName, firstName, lastName, role, identifier, pronoun, hyperonym, hyponym, meronym, synonym (values can concatenate)
- For groups, you can add @corresp and @subtype to indicate the single persons entering the composition of the group; values of subtype: full, partial, potential

Exercise 2: your samples (xml files)

Natalja Tomic, English 1
Vojtech Malinek, Polish 1
Anders Malvik, English 2
Lovro Skopljanac, German 1
Katie Mishler, French 1
Katrin Horn, Italian 1
Lauren Cassidy, English 3
Andrejka Zejn, English 4
Luiza Marinescu, Romanian 1
Fotini Kaidaki, Spanish 1
Jana-Katarina Mende, Polish 2
Susana Sotelo Docio, Spanish 2
Marek Debnar, Polish 3
Ioana Lionte, Romanian 2
Alexandra Olteanu, Romanian 3
Lucretia Pascariu, Romanian 4
Cezary Rosinski, Polish 4
Lucija Mandic, English 5
Lou Burnard, French 2

Exercise 2: tools

- Notepad++, SublimeText, BBEdit, etc.
- Better: oXygen + schema (tei_NER.rng)

Exercise 2: questions

- For the last part of the TS, please prepare an answer to the following questions:
  - How many NE have you annotated?
  - How many per category?
  - Were the proposed values for @ana and @type enough?
  - Did you use @corresp and @subtype?
  - Any ideas about how to interprete your annotations?