Seminar: *Korpusbasierte englische Sprachwissenschaft*
Dr. Elke Teich
Winter Semester 2001/2002

Multilingual Part-of-Speech Tagging

Student: Garance PARIS

8th February 2002

English Linguistics and Translation Studies
Department 4.6 Applied Linguistics, Translation and Interpreting
University of the Saarland

Multilingual PoS-Tagging

To what extent are tag sets for different languages comparable?

What characteristics of a tag set can help guarantee equivalence?

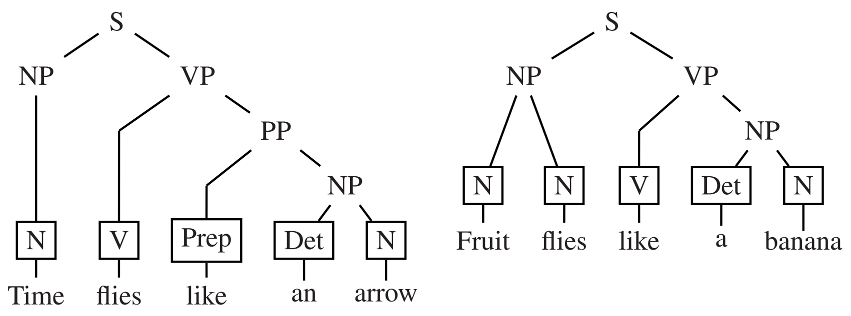
How do differences in the tag sets influence the quality of the information we can extract from a multilingual corpus?

- tag parallel corpus
- compare tag sets
- extract illustratory syntactic constructions

PoS-Tagging?

Def.: Labelling of each word with its appropriate part-of-speech

Broader goal: parse the sentences
Tagging = Specify the preterminal nodes of the parse tree
(→ partial syntactic disambiguation)



(Garside et al., 1997; Manning and Schütze, 1999)

Probabilistic Tagging

Sources of information for the tagger?
(cf. TreeTagger parameter files: Schmid, 1994, 1995; Stein and Schmid, 1995)

- Possible categories for a word
  play: verb in 70 %, noun in 30 % of cases
- Some sequences of tags are more common than others:
  - Probability of noun following article is greater than verb
  - In *a new play*,
    probability of *play* being a noun = 99.6 %,
    probability of *play* being a verb = 0.4 %

Problems for probabilistic taggers:

- unknown words
- sequences not encountered during training

(Non-)Equivalence of Tag Sets

- Multiple tag sets for any one language, which differ in their labels and number of tags
- Different quality measures:
  external validity: choice of tags allows retrieval of grammatical distinctions in subsequent processing
  internal validity: choice of tags increases tagging accuracy
- Differences due to
  1. degree of precision
  2. application (inclusion of some lexical semantics)
  3. tagging method (choice of tags in order to increase accuracy)
  4. divergence of linguistic theories as to extension of PoS-category
  5. irrelevance of some PoS categories for some languages

  (Véronis and Khouri, 1995)

Tag Sets Comparison

- German: Stuttgart-Tübingen tag set (Schiller et al., 1995)
  French: A. Steins tag set (Stein, 1995)
- Hierarchical tag sets, decomposable tags
  E.g. KOUS=**KO**njunktion-**U**nterordnend-**S**atz
  KOUI=**KO**njunktion-**U**nterordnend-**I**nfinitiv
  $\rightarrow$ allows search at different levels of granularity
  ("wildcards": regular expressions)
- Size of tag sets roughly similar: STTS 54, AS 62

Tagging Schemes

None available for French parameters of TreeTagger;
STTS: Schiller et al. (1995)

- Multiwords
  - STTS: technically too difficult to tag as one word
    $\rightarrow$ tag according to syntactic function
    New <NE> York <NE>
    so <ADV> daß <KOUS>
  - AS: only a few are "lexicalised", the majority are broken down into their parts
    alors <CON:sub:1st> que <CON:sub:2nd>
    tel <ADJ> que <CON:sub>

- Mergers
  - STTS: zum <APPRART>
  - AS:
    - du=de+le <PRE:det>
      des=de+les <PRE:det>
    - involved treatment of apostrophes and hyphens
      → hyphenated words not segmented, but well *cette maison –ci*,
      exception: *celle-ci*
      → *l'enfant*, *s'habiller*, but *aujourd'hui*, *d'autant*

Building the corpus

1. Extract 2 months of texts from Le Monde Diplomatique CD
2. "Clean up" data to increase tagging accuracy (perl script):
   - Remove anything that isn't "continuous" text: page numbers, footnotes, keywords, text id numbers...
   - Lowercase all-caps words at the beginning of a line
3. Align with Déjà Vu
   (Units = full sentences, i.e. main clauses containing a finite verb and their dependents)
4. Export "source" and "target" separately as pure text
5. Before tagging, insert SGML tags as segment delimiters
   + number the segments (perl script)
6. Tag with TreeTagger
7. Re-build the parallel corpus in order to have concordances
   (perl-script)

Extracting Information from the Corpus

- Attributive pronouns instead of determiners: extracting NPs still possible by lumping together with articles

```
det   --> <ART>
det   --> <PDAT>
det   --> <PIAT>
det   --> <PPOSAT>
det   --> <PWAT>
```

```
<np><pos ART>Die
  <n1><pos NN>EU-Partner</n1></np>
haben sich darauf verständigt, [...]
```

```
<np><pos PPOSAT>Sein
  <n1><pos NN>Mandat</n1></np>
lautet zusammengefasst:
```


- Lumping together PRELS and PRELAT (both corresponding to PRO:rela) enables extraction of relative clauses

```
rpro --> <PRELS>
rpro --> <PRELAT>
```

```
Damit hält sich die WTO freilich nur an die Grund-
prinzipien,
  <rc><pos PRELS>die
    <np>ihr ganzes Handeln</np>
    <v1><pos VVFIN>bestimmen</v1></rc>
.
```

```
Auch das Elektrizitätsunternehmen Endesa -
  <rc><pos PRELAT>dessen
    <np>Mehrheitsaktionärin</np>
    <np>die spanische Gesellschaft Conama </np>
    <v1><pos VAFIN>ist</v1></rc>
- wird angeklagt: [...]
```

- Ibid. with PWS, PWAT for the extraction of interrogative main clauses

- But: One tag for direct and indirect interrogative "pronouns" (adverbs?) makes it difficult to extract subordinate clauses
  - STTS: <PWAV>
    Wo<PWAV> *wohnt er ?*
    sie fragt, wo<PWAV> *er wohnt*
    der Grund, warum<PWAV>
    (argument in Schiller et al. (1995): same distribution)
  - AS: <PRO:inter>
    où<PRO:inter> *habite-t-il?*
    elle demande où<PRO:inter> *il habite*

- AS's only tag for ADJ makes it impossible to extract only predicative adjectives in French
- Both French and German use present participles as adjectives, but absence of tag in STTS does not easily allow extraction of correspondences
- No annotation for case in STTS → what about correspondence between German cases and French "formes conjointes, disjointes" + clitic pronouns?
- Many "de" in French: how to distinguish those which correspond to PTKZU in German?

Conclusion

- Both tag sets from IMS in Stuttgart
  → same linguistic theory (cf. "attributive" pronouns)
  → less problems (how generalisable is this?)
- This put apart,
  To what extent are tag sets for different languages comparable?
  What characteristics of a tag set can help guarantee equivalence?
  - monolingual comparison: simplifications increasing accuracy (STTS: verbs, adjectives) risk reducing comparability
  - multilingual comparison:
    - One source of differences = relevant categories
      → harmonisation of tag sets should allow for extra categories
    - Mapping betw. tag sets: too much info is better than too little (joining sets possible, not making distinctions after tagging)
      Problem: how to achieve high accuracy with very detailed tagset (given unknown words/sequences)

- *How do differences in the tag sets influence the quality of the information we can extract from a multilingual corpus?*
  - monolingual aspect: tag sets should be really hierarchical
  - multilingual aspects:
    - tokenisation (tagging scheme):
      words for grammatical relations (closed class) → one unit
      (although compensation by query language in part possible)
    - differences between languages and thus between tag sets must of course be taken into account in query
    - here also: the more detailed the annotation, the better
    - sometimes when annotation too coarse, other syntactical information may allow to recover (redundancy), but not always
      e. g. word order in subordinate clauses for German PVAW, but impossible for French PRO:inter
    - difficulties in extracting information are often simply due to differences between languages
      annotation **helps** if queries well formulated

Tagging with the TreeTagger

- Log on to `troff.at.coli.uni-sb.de` (ssh2 only)
  Login: `eteich`
  Password: ?

  (Under Windows in the DI computer lab, click on `Netzwerkumgebung`, `FR86cip`$\boxed{\texttt{sv}}$, `Apps`, `Putty`
  Click on `Connect`, `SSH`, check "2" as your `Preferred SSH protocol version`
  Click on `Session`, enter hostname and choose `ssh`)
- Add the tagger to the path:

```
$ export PATH=$PATH:/export/black/tmp/TreeTagger/cmd
$ export PATH=$PATH:/export/black/tmp/TreeTagger/bin
```

- Tagging:

```
$ echo "Peter sees a horse." | tree-tagger-english
$ echo "Dies ist ein Test." | tree-tagger-german
$ echo "Paul veut bien que Marie vienne." | tree-
tagger-french
```


Querying the corpus with Gsearch

- Use ssh2 to log on to `gnome.coli.uni-sb.de`
  Login: `eteich`
  password: ?
- If necessary, extend the Phrase Structure Grammars in `~/Programs/Gsearch/Grammars`:

```
% Rules:
np --> det adj* noun
np --> pn
% Terminals:
det --> <ART>
```

- Simple queries (syntax and options: see Corley et al., 2001)
  - Searching in the French originals:

```
gsearch -a1 -p pos    concord -         "<lemma=moment>
gsearch -a1 [options] concord [grammar] [goal]
```


  - Searching in the German translations:

```
gsearch -b1 -p pos    concord -         "<lemma=zahllos
gsearch -b1 [options] concord [grammar] [goal]
```

- Complex queries:

```
gsearch -a1 -sc concord - "<lemma=présenter>" > TMP
transform-subcorpus.pl TMP > TMP2
gsearch -b1 -sf TMP2 concord - "<lemma=präsentieren>"
```

```
gsearch -a1 [options] -sc concord [fr-gram] [fr-
goal] > TMP
transform-subcorpus.pl TMP > TMP2
gsearch -b1 [options] -sf TMP2 concord [de-gram] [de-
goal]
```


References

Martin Corley, Steffan Corley, Matt Crocker, Frank Keller, and Shari Trewin. *Gsearch User Manual*, September 28, 2001. Revision 1.3 of the manual for Gsearch version 2.06w, available from `http://www.hcrc.ed.ac.uk/gsearch/src/manual.ps.gz`.

Roger Garside, Geoffrey Leech, and Anthony McEnery. *Corpus Annotation. Linguistic Information from Computer Text Corpora*, chapter 2 and 7. Longman, London, 1997.

Christopher Manning and Hinrich Schütze. *Foundations of statistical and natural language processing*, chapter 10. MIT Press, Cambridge, MA, 1999.

Anne Schiller, Simone Teufel, and Christine Stöckert. Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart. Available from `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/stts_guide.ps.gz`, 1995.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, September 1994. Revised version of the paper available from `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.ps.gz`.

Helmut Schmid. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL SIGDAT-Workshop*, March 1995. Revised version of the paper available from `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger2.ps.gz`.

Achim Stein. Liste der Tags in französichen Korpora. Available from `http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html`, June 1995.

Achim Stein and Helmut Schmid. Étiquetage morphologique de textes français avec un arbre de décisions. *Traitement Automatique des Langues, Numéro spécial: Traitement probabilistes et corpus*, 1-2(36):2336, 1995. Available from `http://www.uni-stuttgart.de/lingrom/stein/pubs/steins95.ps.gz`.

Jean Véronis and Liliane Khouri. Étiquetage grammatical multilinge: modèle. Centre National de la Recherche Scientifique. MULTEXT Project, Document LEX2 (draft), available from `http://www.lpl.univ-aix.fr/projets/multext/LEX/LEX2.html`, 1995.