Linguistically Annotated Learner Corpora: Aspects of a Layered Linguistic Encoding and Standardized Representation

Detmar Meurers and Holger Wunsch

Seminar für Sprachwissenschaft, Universität Tübingen
{dm,wunsch}@sfs.uni-tuebingen.de

Motivation

Corpus data can be a **valuable source of empirical insight** for building and falsifying linguistic theories.

- **Linguistic annotation** is crucial for their **effective exploration** from a linguistic perspective.
- Annotation needs to be consistent to be useful.

For studying the properties of language learners, learner corpora are starting to become available (Granger, 2008)

- Annotation generally focused on marking errors.
- Whether and how error annotation with high interrater agreement can be obtained is unclear.

Second Language Acquisition (SLA) research is concerned with identifying linguistic regularities in learner language (e.g., Pienemann, 1998), independent of whether it is well-formed or not.

➡ For effective exploration of learner corpora, it is crucial to annotate them linguistically.

➡ What linguistic categories are adequate and consistently applicable for the interlanguage of language learners?

POS-tagging learner language

POS taggers differ in the methods and fallback strategies they use, e.g.:

- Suffix analysis and other morphological clues
- Exploit evidence from distribution

Learner language typically contains a large number of forms that do not occur in native language, i.e., unseen forms requiring fallback strategies.

We POS-tagged the **NOCE corpus** with three taggers: TnT (Brants 2000), Stanford Tagger (Toutanova & Manning 2000), TreeTagger (Schmid 1994).

- NOCE is an error-annotated corpus of essays written by Spanish learners of English (Díaz-Negrillo 2007).
- All taggers were trained on the same training data (Wall Street Journal)

➡ Do differences in the POS assignment of the three taggers point to erroneous forms in learners' writings?

1. Manual inspection

(1) *[…] it has **grew** up a lot specially after 1996 […]*

- TreeTagger: **past participle (VBN)**
- Stanford Tagger: **past tense verb (VBD)**
- TnT: **past tense verb (VBD)**

➡ Verb form error (distribution-morphology mismatch, see (9))

2. Quantitvative assessment

Can diverging POS assignments be used to find errors **automatically**?

- Look for words with diverging POS tags.
- Check whether this word was marked up as being erroneous.
- Only look at errors spanning one word, ignore other errors.

	Different	Equal	
Error	731	2558	**Precision:** 12.1 %
No error	5287	34927	**Recall:** 22.2 %

➡ Diverging POS assignment correlates with errors only to a low degree.

➡ Instead of targeting errors, develop linguistic annotation scheme.

References

Brants, T. (2000). TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*. pp. 224–231.

Díaz-Negrillo, A. (2007). *A Fine-Grained Error Tagger for Learner Corpora*. Ph.D. thesis, University of Jaén, Spain. Revised version published 2009 as *EARS: A User's Manual*. Munich, Germany: LINCOM Academic Reference Books.

Díaz-Negrillo, A., D. Meurers, S. Valera and H. Wunsch (2010). Towards Inter-language POS Annotation for Effective Learner Corpora in SLA and FLT. *Language Forum*. In press. http://purl.org/dm/papers/diaz-negrillo-et-al-09.html

Granger S. (2008) Learner Corpora. In Lüdeling A. and M. Kytö (eds) *Corpus linguistics: an international handbook, vol. 1*. Mouton de Gruyter. pp. 259–275.

Pienemann, M. (1998). *Language Processing and Second Language Development: Processability Theory*. John Benjamins.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK.

Toutanova, K. and C. Manning (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on EMNLP and VLC*. pp. 63–70.

Part-of-speech analysis revisited

POS tagging is a disambiguation task considering **multiple sources of evidence** (linguistic layers).

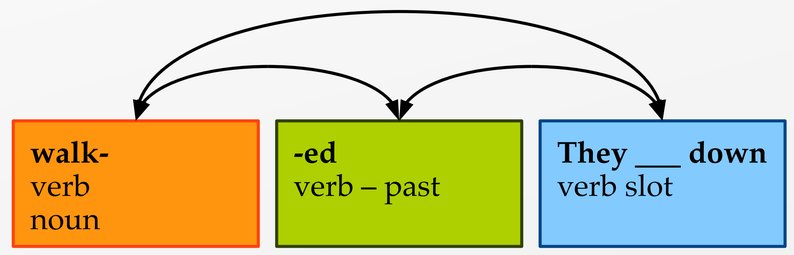
(2) *I was surprised by the word **of** the day.* — **Lexis:** Preposition (of)

(3) *His son **brachiated** along the monkey bars.* — **Morphology:** Verb (suffix *-ed*)

(4) *They **man** the boat.* — **Distribution:** Verb (*man* located in verb position)

➡ Usually, the sources of evidence **converge** on one classification:

(5) *They **walked** down the street.*



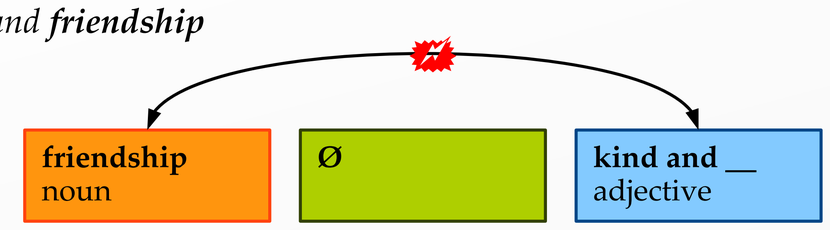
Layered POS categories for learner language

Errors in POS-assignment occur where the **linguistic sources of evidence diverge** (see also Díaz-Negrillo, Meurers, Valera, Wunsch 2009).

Prototypical cases of divergence of evidence with POS tagging

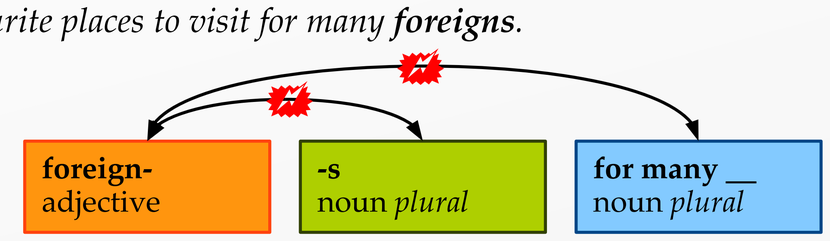
1. Stem-distribution mismatch

(6) *They are very kind and **friendship***



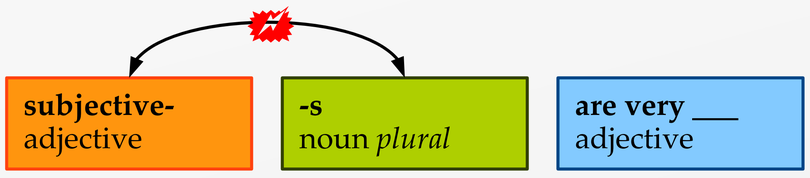
2. Stem-distribution, stem-morphology mismatch

(7) *[…] one of the favourite places to visit for many **foreigns**.*



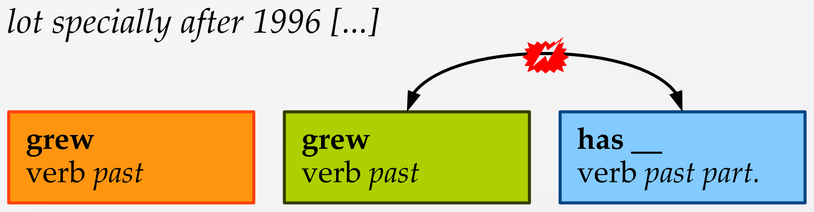
3. Stem-morphology mismatch

(8) *[…] television, radio are very **subjectives** […]*



4. Distribution-morphology mismatch

(9) *[…] it has **grew** up a lot specially after 1996 [...]*



Conclusions and future work

Errors in learner language can be analyzed as an epi-phenomenon resulting from conflicting properties on different linguistic layers.

For Part-of-speech tagging, we identified three relevant sources of evidence: lexis, morphology, and distribution.

While for native language, these sources of evidence usually converge, we showed that this is not necessarily the case for learner language.

➡ We devise a tri-partite system of POS categories which can consistently represent both diverging and converging sources of evidence.

Directions for future work:

- Implementation of a tri-partite POS tagger for learner language.
- Generalization of the *n*-partite model of categories to other domains. For example, the syntactic property of constituency can be described in terms of the linguistic levels of dependency, topology, and precedence.