On emergent linguistic characteristics in learner and translation corpora

Detmar Meurers
Universität Tübingen

Université Paris 7, UFR Études Interculturelles de Langues Appliquées (EILA)
March 21, 2011

Overview

- Context: A sketch of our research perspective
- On linguistically analyzing learner language
  - Categories for interlanguage
    - Parts-of-speech as an example:
      – sources of evidence
      – nature of categories
    - Which level of analysis?
      – between robustness and representing variation
  - Target hypotheses and error annotation
    - Inter-annotator agreement and gold-standards
    - Comparative fallacy
  - Relevance of the task and learner modeling
- Emerging, data-driven units in translation corpora?
  - Automatically detecting variation in corpus annotation to detect annotation errors (DECCA)
  - Considering a related approach for translation corpora

On emergent linguistic characteristics in learner and translation corpora
Detmar Meurers
Introduction
Background
Learner Corpora
Data in SLA Research
Corpus annotation
Categories for Learner Language
Example: Parts-of-speech
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence
Systematic categories
Nature of interlang. categories
Comparative fallacy
Syntactic annotation
Error annotation
Target hypotheses
Activity & learner modeling
Task-specific learner corpora
Emerging Units in Translation Corpora
Variation detection
Issues in working out the idea
Conclusion

Context: A sketch of our research perspective

Analyzing learner language

- Intelligent Tutoring System TAGARELA for Portuguese (Amaral & Meurers 2008, 2009, 2011; Amaral et al. 2011)
- Automatic analysis of learner language (Meurers 2009)
- Linguistic analysis of NOCE corpus of English written by Spanish learners (Díaz-Negrillo, Meurers, Valera & Wunsch 2010)
- Word order errors (Metcalf & Meurers 2006b; Boyd & Meurers 2008)
- Content assessment of answers to reading comprehension questions (Bailey & Meurers 2008) → CoMiC (SFB 833 A4)
  - Longitudinal corpus collection using WELCOME (Meurers, Ott & Ziai 2010a) → KU/OSU collaboration
  - Dependency parsing of learner language (Ott & Ziai 2010)

Analyzing language for learners

- Visual input enhancement of authentic web pages for learners (WERTi, Meurers et al. 2010b)
- Language-aware search engine (Ott & Meurers 2010)

On emergent linguistic characteristics in learner and translation corpora
Detmar Meurers
Introduction
Background
Learner Corpora
Data in SLA Research
Corpus annotation
Categories for Learner Language
Example: Parts-of-speech
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence
Systematic categories
Nature of interlang. categories
Comparative fallacy
Syntactic annotation
Error annotation
Target hypotheses
Activity & learner modeling
Task-specific learner corpora
Emerging Units in Translation Corpora
Variation detection
Issues in working out the idea
Conclusion

Contact Points: CL & learner language analysis

- **Learner corpora**: representing, annotating, searching
  - can provide empirical evidence for SLA research
  - can provide insights into typical student needs in FLT

  annotation = off-line analysis
- **Writer's aid tools**: on-line analysis of learner language to provide immediate feedback *aimed at producing text*
- **Language testing**: off-line or on-line analysis to support or automate *assessment of learner abilities*
- **Intelligent Tutoring Systems**: on-line analysis
  - to provide immediate, individualized feedback, e.g.:
    - meta-linguistic feedback in a form-focused activity
    - incidental focus-on-form in a meaning-based activity
    - feedback on meaning (very rare in ITS)
  - to determine progression through pedagogical material

  aimed at supporting language acquisition.

On emergent linguistic characteristics in learner and translation corpora
Detmar Meurers
Introduction
Background
Learner Corpora
Data in SLA Research
Corpus annotation
Categories for Learner Language
Example: Parts-of-speech
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence
Systematic categories
Nature of interlang. categories
Comparative fallacy
Syntactic annotation
Error annotation
Target hypotheses
Activity & learner modeling
Task-specific learner corpora
Emerging Units in Translation Corpora
Variation detection
Issues in working out the idea
Conclusion

Data in SLA research

An example: Clahsen & Muysken (1986)

- They studied word order acquisition in German by native speakers of Romance languages.
- Stages of acquisition:
  1. S (Aux) V O
  2. (AdvP/PP) S (Aux) V O
  3. S V[+fin] O V[-fin]
  4. XP V[+fin] S O
  5. S V[+fin] (Adv) O
  6. *dass* S O V[+fin]

  Stage 2 example: *Früher* *ich kannte* *den Mann*
  earlier$_{AdvP}$ I$_S$ knew$_V$ [the man]$_O$

  Stage 4 example: *Früher* *kannte* *ich den Mann*
  earlier$_{AdvP}$ knew$_{V[+fin]}$ I$_S$ [the man]$_O$
- How is the data characterized?
  - lexical and syntactic categories and functions
  - some acquisition stages are well-formed, others ill-formed

Annotation: Error Annotation and Beyond

- SLA research essentially observes correlations of linguistic properties, whether erroneous or not.
- Yet, the annotation of learner corpora has focused on *errors* made by the learners (cf. Granger 2003; Díaz-Negrillo & Fernández-Domínguez 2006).
- Even where errors are the research focus, their correlation with other linguistic properties is relevant.
- A wide range of linguistic modeling useful for capturing
  - overuse/underuse of particular patterns
  - measures of language development
    - CAF (Wolfe-Quintero et al. 1998; Ortega 2003; Housen & Kuiken 2009; Lu 2010)
    - Criterial Features (Hawkins & Buttery 2009, 2010)

Annotation of Linguistic Properties

- Annotation schemes for native language corpora have been developed for a wide range of linguistic properties, including
  - part-of-speech and morphology
  - syntactic constituency or lexical dependency structures
  - semantics (word senses, coreference), discourse structure
- Each type of annotation typically requires an extensive manual annotation effort → gold standard corpora
- Automatic annotation tools learning from such gold standard annotation are becoming available, but
  - quality of automatic annotation drops significantly for text differing from the gold standard training material
- Interdisciplinary collaboration between SLA & CL crucial to adapt annotation schemes & methods to learner language
  - Surprisingly little research on this (Meunier 1998; de Haan 2000; de Mönnink 2000; van Rooy & Schäfer 2002, 2003).

Annotation Schemes and Annotation Quality

- An annotation scheme is only as good as the distinctions it reliably supports making based on available evidence.
  - E.g., particle vs. preposition dropped in Penn Treebank tagset since often not enough evidence available.
  - Note: More classes may be more reliable if they are more coherent (cf. CLAWS7 annotation, followed by mapping to CLAWS5 in BNC Tag Enhancement Project).
- How can high quality gold standards be obtained?
  - Keep only reliably and consistently identifiable distinctions, described in detailed manual, including appendix on hard cases (Voutilainen & Järvinen 1995; Sampson & Babarczy 2003)
  - Annotate corpus several times and independently, then test interannotator agreement (Brants & Skut 1998)
  - Detection of annotation errors through automatic analysis of comparable data recurring in the corpus → DECCA (Dickinson & Meurers 2003a,b, 2005b; Boyd et al. 2008)

Linguistically annotating learner language

Parts-of-speech as an example

- The NOCE learner corpus (Díaz-Negrillo 2007, 2009)
  - Short essays written by Spanish 1st and 2nd year students of English, annotated with editing and error tags
  - 998 texts, 337.332 tokens (149.256 types)

⇒ How about adding linguistic information?
(Díaz-Negrillo, Meurers, Valera & Wunsch 2010)

- Exploring automatic POS annotation
- What does it mean to POS-annotate learner language?

Automatic POS-Tagging of NOCE

- Used 3 POS taggers trained on WSJ newspaper text, using Penn Treebank tagset (TreeTagger, TnT, Stanford)
- Manually evaluated POS tags assigned by taggers to 10 texts by 10 different participants (1.850 words)
  - TreeTagger: 94.95%
  - TnT Tagger: 94.03%
  - Stanford Tagger: 88.11%

Aspects of a qualitative analysis

- We found lower performance for expressions which do not exist in English (cf. also de Haan 2000; van Rooy & Schäfer 2002).

(1) *I think that university teachs to people ...* [spelling]

(2) *They can't pay their studies and more over they have to pay a flat ...* [tokenization]

- But is tagging learner language really just a robustness issue, like adapting taggers to another domain?
- What does it mean to use POS tags developed for native language for the interlanguage of learners?
  - What research questions can such POS tags answer?

Three Sources of Evidence for POS analysis

Lemma/Lexical entry: *of* ⇒ preposition

(3) *drugs can be killer **of** many of ours.*

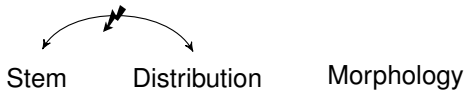
Morphology: *-ion* ⇒ noun

(4) *but it was a **revolution** in that period*

Distribution: det __ noun ⇒ adjective

(5) *In the **modern** life the people can communicate*

Case 1: Stem-Distribution mismatch



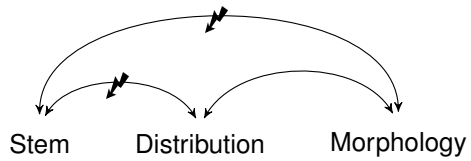
(6) […] *you can find a big **vary** of beautiful beaches* […]

Stem	Distribution	Morphology
verb	noun	?

(7) *RED helped him **during** he was in the prison.*

Stem	Distribution	Morphology
preposition	conjunction	?

Case 2: Stem-Distrib./Stem-Morph. mismatch



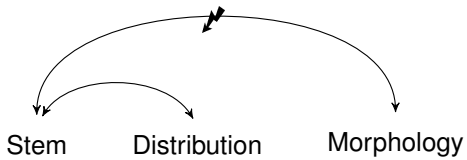
(8) […] *one of the favourite places to visit for many **foreigns**.*

Stem	Distribution	Morphology
adjective	noun	noun / verb 3rd sg

(9) […] *to be **choiced** for a job* […]

Stem	Distribution	Morphology
noun / adjective	verb	verb

Case 3: Stem-Morphology mismatch



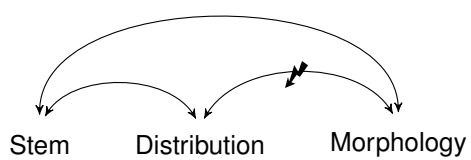
(10) […] *this film is one of the **bests** ever* […]

Stem	Distribution	Morphology
adjective (noun / verb)	adjective	noun / verb 3rd sg

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

Stem	Distribution	Morphology
adjective / noun	adjective	noun / verb 3rd sg

Case 4: Distribution-Morphology mismatch



(12) […] *for almost every **jobs** nowadays* […]

Stem	Distribution	Morphology
noun	noun sg	noun pl / verb 3rd sg

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

Stem	Distribution	Morphology
verb	verb past participle	verb past tense

Systematic POS for Learner Language

- A single, standard POS tag **fails to systematically identify properties of learner language**.
- Alternative: tripartite POS encoding of
  - distribution, stem, morphology
- Some errors in learner language are **epiphenomena of mismatches** in linguistic encoding.
  - → Identify such errors through linguistic annotation.
- The value of identifying such mismatches systematically is confirmed by recent SLA research (Zyzik & Azevedo 2009)
  - L2 learners have difficulty distinguishing between word classes among semantically related forms
  - Hypothesis: L2 learners have limited ability to interpret syntactic and morphological cues!

On the nature of categories for learner language

- Where do the categories abstracted to come from?
- Categories result from generalizations, which require a significant amount of comparable data to be made.
  - requires decision on what constitutes comparable data, which is difficult for a dynamic target such as interlanguage
- Robustness and the level of analysis:
  - In NLP, *robustness* is the ability to *ignore* variation in the realization of a category to be identified.
  - But variation in the realization of a category is an important characteristic of learner language.
    - → Design annotation schemes for learner language to encode minimal observations.
    - → Provide access to those on one level of annotation, with other annotation levels providing robust L2 abstractions.

On the nature of categories for learner language

Comparative fallacy

- "mistake of studying the systematic character of one language by comparing it to another" (Bley-Vroman 1983)
  - extended to include bias towards towards native language (Lakshmanan & Selinker 2001)
- Essentially trying to analyze a "non-canonical variety" using a "robust" version of the canonical grammar.
  - divergences from norm is annotated as errors
  - but: the research question is the issue here, not corpus error annotation as such (Tenfjord et al. 2006)
- Issue more general than language acquisition research:
  - Eurocentrism in field work, e.g., Gil (2001)
  - Variationist sociolinguistics

→ Importance of explicitly defining classes and when an instance is counted as one of the variants.

On the nature of categories for learner language

Aspects of syntactic modeling

- Just like POS categories, syntactic structure is motivated by different types of evidence.
- For analyzing learner language, one can separate:
  - overall topology of a sentence (Hirschmann et al. 2007)
  - chunks and chunk-internal word order (Abney 1997)
  - lexical dependencies
    - canonical, as interface to meaning (MacWhinney 2008; Rosén & Smedt 2010; Ott & Ziai 2010; Hirschmann et al. 2010)
    - non-canonical, separating evidence for morpho-syntactic and semantic relations (Dickinson & Ragheb 2009)

Error annotation

- Error annotation involves (implicitly or explicitly):
  a) Determining what the learner wanted to say (**target**).
  b) Identifying
     i. the **location** of the error, and
     ii. the **type** of the error corresponding to the difference between the learner sentence and the target hypothesis.
  c) **Annotating** the error in the corpus
- Each of these steps can present ambiguity:
  a) multiple possible target hypotheses
  b) i. different locations in which the error can be rooted
     ii. different types of errors a divergence can be attributed to
  c) different ways to mark an error location & type in corpus

Error annotation schemes: Desiderata

Inter-annotator agreement

- An annotation is only relevant and useful if it provides a uniform, reliable index to relevant classes of data.
- Traditionally every researcher develops their own error annotation scheme. (cf. Díaz-Negrillo & Fernández-Domínguez 2006)
- Alarmingly, no studies on which inter-annotator agreement can be reached for which distinctions in error annotation
- No freely available gold standard corpora, so
  - no reliable quantitative evaluation in research
  - no reliable training & evaluation of NLP for error analysis
- Promising progress for some subclasses (det, prep)
  (e.g., Lee & Seneff 2006; Tetreault & Chodorow 2008; De Felice 2008)
  - but it is important to establish a tool-independent, transparent definition of the markables to be annotated

Target hypotheses

- Fitzpatrick & Seegmiller (2004) report unsatisfactory levels of agreement in determining learner target forms.
  - Keeping the target hypothesis implicit results in error annotation which diverge even more unsatisfactorily.
- Anke Lüdeling has argued for making target hypotheses an explicit part of error annotation (Lüdeling et al. 2005; Hirschmann et al. 2007; Lüdeling 2008).
  - supports alternative targets (and corresponding error annotation), and
  - supports multiple levels of target hypotheses, differing in scope and operations allowed to obtain them
    - e.g., only replacement, omission, etc. to make sentence locally well-formed vs. taking context into account
- If target hypothesis is explicit, one can evaluate reliability of second step, from target hypothesis to error tag.

Difficulty of determining target hypotheses

- What are the target forms for the sentences taken from the Hiroshima English Learners' Corpus (Miura 1998)?

  (14) *I didn't know*
  (15) *I don't know his lives.*
  (16) *I know where he lives.*
  (17) *I know he lived*

  They are taken from a translation task, for the Japanese of

  (18) *I don't know where he lives.*

- How can one obtain a better handle on target hypotheses?
  - Focus on more advanced learners.
  - Take explicit task context into account.
  - Support targets other than fully explicit surface forms.
  - Take more learner strategies into account.
    - Learners often lift material from texts or use mastered chunks instead of trying to express appropriate meaning!

Constraining the search space of interpretation

Importance of activity and learner modeling

- All approaches to modeling learner language, such as
  - *mal*-rules, constraint relaxation, statistical modeling

  must model the space of **well-formed and ill-formed variation** that is possible given
  - a particular activity, and
  - a given learner.
- For example, without task and speaker context, how would you interpret the following?

(19) *I will not buy this record it is scratched*

(20) *My hovercraft is full of eels.*

Exemplifying interpretation in context

Monty Python: Hungarian Phrase Book sketch
http://purl.org/net/mp-sketch

Towards task-specific learner corpora

- Explicit task and learner models included as meta-information in a corpus can provide crucial constraining information for interpreting learner language.
  - E.g., it's easier to infer what a learner wanted to say if one knows the text they are answering questions about.
  - Related to taking strategic competence, task, and L1 into account in learner models of Intelligent Tutoring Systems (Amaral & Meurers 2008).
- Most current learner language corpora consist of essays, yet learners produce language in a wide range of contexts, naturalistic or instructed, e.g.,
  - email and chat messages
  - answering reading or listening comprehension questions
  - asking questions in information gap activities
- To obtain learner corpora which are interpretable and representative, we need language resulting from explicit tasks, in a variety of contexts, including longitudinal data.

Emerging Units in Translation Corpora

- What constitutes relevant linguistic units of analysis?
- How about units for analyzing translation corpora?
- Starting point: Variation *n*-gram error detection approach
  - part-of-speech annotation (Dickinson & Meurers 2003a)
  - syntactic annotation (Dickinson & Meurers 2003b; Boyd, Dickinson & Meurers 2007)
  - discontinuous syntactic annotation (Dickinson & Meurers 2005b)
  - dependency annotation (Boyd, Dickinson & Meurers 2008)
  - spoken language corpora (Dickinson & Meurers 2005a).
- Idea: Use the approach to study the variation in recurrent *n*-grams in translation corpora
- Work just started, so we here discuss issues arising in working out idea, as basis for feedback and discussion.

Variation Detection for POS Annotation

(Dickinson & Meurers 2003a)

- POS tagging reduces the set of lexically possible tags to the correct tag for a specific corpus occurrence.
  - A word occurring multiple times in a corpus can occur with more than one annotation.
- Variation: material occurs multiple times in corpus with different annotations
- Variation can result from
  - genuine ambiguity
  - inconsistent, erroneous tagging
- How can one find such variation and decide whether it's an ambiguity or error?

Classifying variation

- The key to classifying variation lies in the context:
  - The more similar the context of the occurrences, the more likely the variation is an error.
- A simple way of making "similarity of context" concrete is to say it consists of
  - words
  - which immediately surround the variation, and
  - require identity of contexts.

⇒ Extract all *n*-grams containing a token that is annotated differently in another occurrence of the *n*-gram in corpus.

- variation nucleus: recurring unit with different annotation
- variation *n*-gram: variation nucleus with identical context

Computing variation *n*-grams

- Example from WSJ: Variation 12-gram with *off*

(21) *to ward off a hostile takeover attempt by two European shipping concerns*

- once annotated as a preposition (IN), and
- once as a particle (RP).

- Note: Such a 12-gram contains two variation 11-grams:

(22) *to ward off a hostile takeover attempt by two Eur. shipping*
ward off a hostile takeover attempt by two Eur. shipping concerns

→ Calculate variation *n*-grams based on variation *n*–1-grams to obtain an algorithm efficient enough for large corpora.

- Essentially an instance of the a priori algorithm (Agrawal & Srikant 1994).

Computing variation *n*-grams

Algorithm

1. Calculate the set of variation unigrams in the corpus and store them.
2. Extend the *n*-grams by one word to either side. For each resulting $(n+1)$-gram
   - check whether it has another instance in the corpus and
   - store it in case there is a variation in the way the occurrences are tagged.
3. Repeat step 2 until we reach an *n* for which no variation *n*-grams are in corpus.

Running this algorithm on the Penn Treebank 3 version of the WSJ, retrieves variation *n*-grams up to length 224.

Applying the idea to translation corpora

- Idea: Use the same approach to identify translation variation *n*-grams in aligned corpora.
  - View translation as a form of annotation
- First step: Identify recurring units of any length.
- For example, we extracted recurrent n-grams from fr-en Europarl v6 (http://www.statmt.org/europarl)
  - over 47 million tokens (in French part)
  - recurrent n-grams found (length ≥ 2, recurrence ≥ 2):
    - longest: 621 tokens
    - total number: 18.181.667

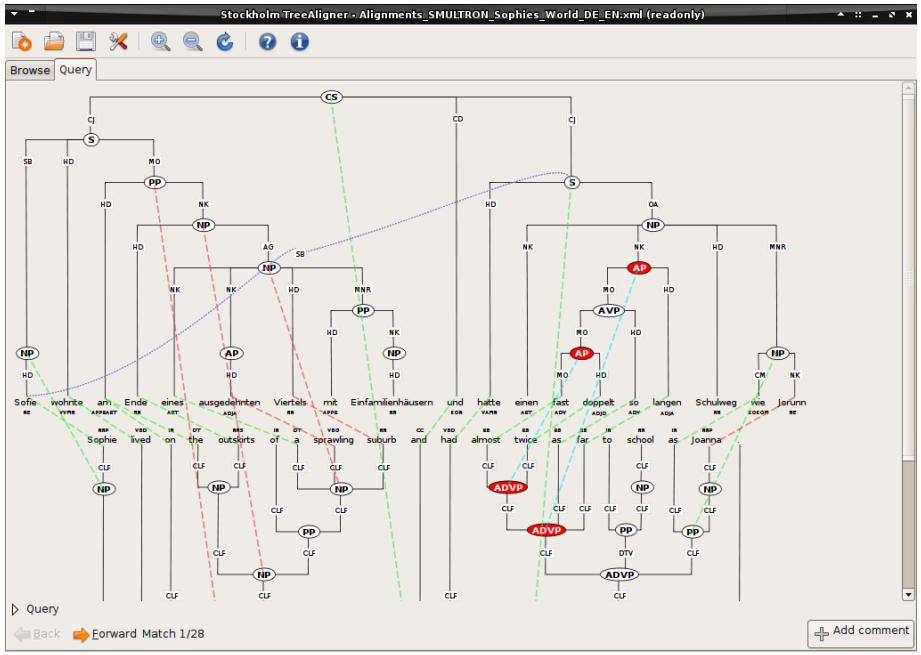
Which recurring units are relevant?

- Problem: Every n-gram contains two n-1 grams.
- For example:
  - *conclusion d'un protocole portant* (4, 5)
  - *d'un protocole portant adaptation* (4, 5)
  - *protocole portant adaptation des* (4, 5)
  - *portant adaptation des aspects* (4, 5)
- Flood of n-grams not interesting for unit identification.
- More interesting units are needed!
- Potential solution: Keep only those recurrent n-grams which cannot be further extended.
  - For every token in the corpus, record it only as part of the longest recurring n-gram type that it is apart of.

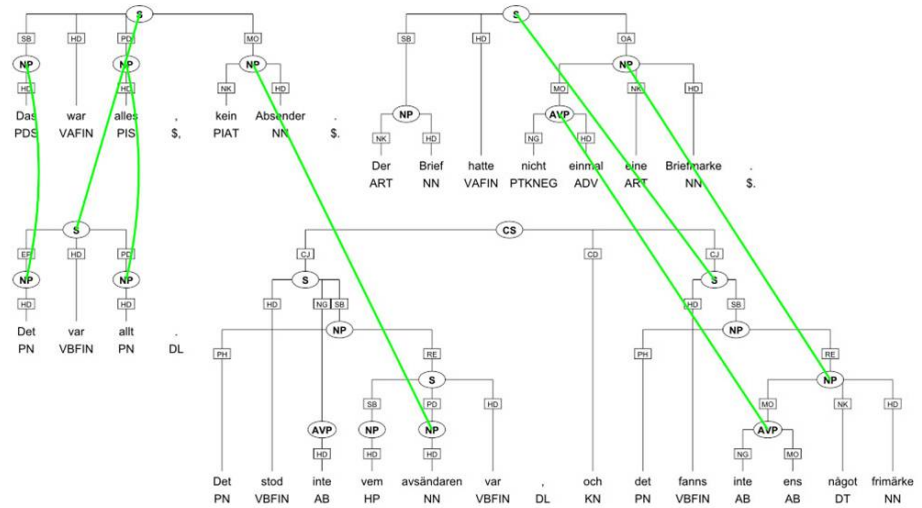
On identifying translation variation

- When viewing translation as annotation, we need to consider which units are aligned in the corpus.
- Europarl is sentence aligned, but we want to look at the translation of recurring n-grams, which can be smaller or bigger than sentences.
- What are useful ways to characterize variation in translation for a corpus which (only) is sentence aligned?
  a) Compare length of longest recurring unit in translation of corresponding sentence.
     - Problematic for sentences with multiple recurring n-grams
  b) Only consider recurring sentences (not any n-grams).
     - limits method to 4372 recurrent cases in Europarl-fr
     - many seem to arise through genre, not language:
       – *Le procès-verbal de la séance d'hier a été distribué. Y a-t-il des observations ?* (2, 44)
       – *La discussion commune est close. Le vote aura lieu à 17h30.* (2, 44)
  c) Alternative: Use a corpus with a richer correspondence → aligned treebanks, e.g., SMULTRON (Volk et al. 2010)

SMULTRON example



Domains beyond sentences: *m*-to-*n* tree alignment



Summary of automatic translation variation detection idea

- potentially fruitful to study variation in translation using variation-n-gram detection method
- requires synchronization of those units which can be detected as recurring & those which have been translated
- sketched some potential ideas which we intend to explore (studying statistical MT literature may well provide more)

Conclusion

- We motivated linguistic annotation to support effective querying for SLA patterns and discussed an approach to the POS analysis of learner language separating
  - lexical, morphological, and distributional information.

  Goal: Corpus annotation systematically characterizing language (native-like as well as learner innovations).
- Turning to error annotation, we argued for inter-annotator agreement as crucial for establishing which distinctions are replicable based on the available information.
- We explored the nature of target hypotheses and argued for explicit task and learner modeling to constrain the search space of interpretation.
- Turning to the question of emerging units, we sketched the use of the variation n-gram method for the identification of recurring units and variation in aligned corpora.

References

Abney, S. (1997). Partial Parsing via Finite-State Cascades. *Natural Language Engineering* 2, 337–344. URL http://www.vinartus.net/spa/97a.pdf.

Agrawal, R. & R. Srikant (1994). Fast Algorithms for Mining Association Rules in Large Databases. In J. B. Bocca, M. Jarke & C. Zaniolo (eds.), *VLDB '94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*. Morgan Kaufmann, pp. 487–499.

Amaral, L., V. Metcalf & D. Meurers (2006). Language Awareness through Re-use of NLP Technology. Pre-conference Workshop on NLP in CALL – Computational and Linguistic Challenges. CALICO 2006. May 17, 2006. University of Hawaii. URL http://purl.org/net/icall/handouts/calico06-amaral-metcalf-meurers.pdf.

Amaral, L. & D. Meurers (2008). From Recording Linguistic Competence to Supporting Inferences about Language Acquisition in Context: Extending the Conceptualization of Student Models for Intelligent Computer-Assisted Language Learning. *Computer-Assisted Language Learning* 21(4), 323–338. URL http://purl.org/dm/papers/amaral-meurers-call08.html.

Amaral, L. & D. Meurers (2009). Little Things With Big Effects: On the Identification and Interpretation of Tokens for Error Diagnosis in ICALL. *CALICO Journal* 27(1). URL http://purl.org/dm/papers/amaral-meurers-09.html.

Amaral, L. & D. Meurers (2011). On Using Intelligent Computer-Assisted Language Learning in Real-Life Foreign Language Teaching and Learning. *ReCALL* 23(1), 4–24. URL http://purl.org/dm/papers/amaral-meurers-10.html.

Amaral, L., D. Meurers & R. Ziai (2011). Analyzing Learner Language: Towards A Flexible NLP Architecture for Intelligent Language Tutors. *Computer-Assisted Language Learning* 24(1), 1–16. URL http://purl.org/dm/papers/amaral-meurers-ziai-10.html.

Bailey, S. & D. Meurers (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In J. Tetreault, J. Burstein & R. D. Felice (eds.), *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*. Columbus, Ohio, pp. 107–115. URL http://aclweb.org/anthology/W08-0913.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning* 33(1), 1–17. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-1770.1983.tb00983.x/pdf.

Boyd, A., M. Dickinson & D. Meurers (2007). Increasing the Recall of Corpus Annotation Error Detection. In *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT-07)*. Bergen, Norway. URL http://purl.org/dm/papers/boyd-et-al-07b.html.

Boyd, A., M. Dickinson & D. Meurers (2008). On Detecting Errors in Dependency Treebanks. *Research on Language and Computation* 6(2), 113–137. URL http://purl.org/dm/papers/boyd-et-al-08.html.

Boyd, A. & D. Meurers (2008). On Diagnosing Word Order Errors. Poster presented at the CALICO Pre-Conference Workshop on Automatic Analysis of Learner Language. URL http://purl.org/net/calico-workshop-abstracts.html#6.

Brants, T. & W. Skut (1998). Automation of Treebank Annotation. In *Proceedings of New Methods in Language Processing*. Sydney, Australia. URL http://wing.comp.nus.edu.sg/acl/W/W98/W98-1207.pdf.

Clahsen, H. & P. Muysken (1986). The availability of Universal Grammar to adult and child learners: A study of the acquisition of German word order. *Second Language Acquisition* 2, 93–19. URL http://slr.sagepub.com/cgi/reprint/2/2/93.pdf.

De Felice, R. (2008). Automatic Error Detection in Non-native English. Ph.D. thesis, St Catherine's College, University of Oxford.

de Haan, P. (2000). Tagging non-native English with the TOSCA-ICLE tagger. In Mair & Hundt (2000), pp. 69–79.

de Mönnink, I. (2000). Parsing a learner corpus. In Mair & Hundt (2000), pp. 81–90.

Dickinson, M. & W. D. Meurers (2003a). Detecting Errors in Part-of-Speech Annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*. Budapest, Hungary, pp. 107–114. URL http://purl.org/dm/papers/dickinson-meurers-03.html.

Dickinson, M. & W. D. Meurers (2003b). Detecting Inconsistencies in Treebanks. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT-03)*. Växjö, Sweden, pp. 45–56. URL http://purl.org/dm/papers/dickinson-meurers-tlt03.html.

Dickinson, M. & W. D. Meurers (2005a). Detecting Annotation Errors in Spoken Language Corpora. In *The Special Session on treebanks for spoken language and discourse at NODALIDA-05*. Joensuu, Finland. URL http://purl.org/~dm/papers/dickinson-meurers-nodalida05.html.

Dickinson, M. & W. D. Meurers (2005b). Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. pp. 322–329. URL http://aclweb.org/anthology/P05-1040.

On emergent linguistic characteristics in learner and translation corpora
Detmar Meurers
Introduction
Background
Learner Corpora
Data in SLA Research
Corpus annotation
Categories for Learner Language
Example: Parts-of-speech
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence
Systematic categories
Nature of interlang. categories
Comparative fallacy
Syntactic annotation
Error annotation
Target hypotheses
Activity & learner modeling
Task-specific learner corpora
Emerging Units in Translation Corpora
Variation detection
Issues in working out the idea
Conclusion
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
39 / 39

Dickinson, M. & M. Ragheb (2009). Dependency Annotation for Learner Corpora. In *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*. Milan, Italy. URL http://jones.ling.indiana.edu/~mdickinson/papers/dickinson-ragheb09.html.

Díaz-Negrillo, A. (2007). A Fine-Grained Error Tagger for Learner Corpora. Ph.D. thesis, University of Jaén, Spain.

Díaz-Negrillo, A. (2009). *EARS: A User's Manual*. Munich, Germany: LINCOM Academic Reference Books.

Díaz-Negrillo, A. & J. Fernández-Domínguez (2006). Error Tagging Systems for Learner Corpora. *Revista Española de Lingüística Aplicada (RESLA)* 19, 83–102. URL http://dialnet.unirioja.es/servlet/fichero_articulo?codigo=2198610&orden=72810.

Díaz-Negrillo, A., D. Meurers, S. Valera & H. Wunsch (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum* 36(1–2). URL http://purl.org/dm/papers/diaz-negrillo-et-al-09.html. Special Issue on Corpus Linguistics for Teaching and Learning. In Honour of John Sinclair.

Fitzpatrick, E. & M. S. Seegmiller (2004). The Montclair electronic language database project. In U. Connor & T. Upton (eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*, Amsterdam: Rodopi. URL http://chss.montclair.edu/linguistics/MELD/rodopipaper.pdf.

Gil, D. (2001). Escaping Eurocentrism: Fieldwork as a Process of Unlearning. In P. Newman & M. Ratliff (eds.), *Linguistic Fieldwork*, Cambridge University Press, pp. 102–132.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal* 20(3), 465–480. URL http://purl.org/calico/granger03.pdf.

On emergent linguistic characteristics in learner and translation corpora
Detmar Meurers
Introduction
Background
Learner Corpora
Data in SLA Research
Corpus annotation
Categories for Learner Language
Example: Parts-of-speech
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence
Systematic categories
Nature of interlang. categories
Comparative fallacy
Syntactic annotation
Error annotation
Target hypotheses
Activity & learner modeling
Task-specific learner corpora
Emerging Units in Translation Corpora
Variation detection
Issues in working out the idea
Conclusion
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
39 / 39

Hawkins, J. A. & P. Buttery (2009). Using Learner Language from Corpora to Profile Levels of Proficiency – Insights from the English Profile Programme. In *Studies in Language Testing: The Social and Educational Impact of Language Assessment*, Cambridge: Cambridge University Press.

Hawkins, J. A. & P. Buttery (2010). Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal* .

Hirschmann, H., S. Doolittle & A. Lüdeling (2007). Syntactic annotation of non-canonical linguistic structures. In *Proceedings of Corpus Linguistics 2007*. Birmingham. URL http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/neu2/mitarbeiter-innen/anke/pdf/HirschmannDoolittleLuedelingCL2007.pdf.

Hirschmann, H., A. Lüdeling, I. Rehbein, M. Reznicek & A. Zeldes (2010). Syntactic Overuse and Underuse: A Study of a Parsed Learner Corpus and its Target Hypothesis. Presentation given at the Treebanks and Linguistic Theory Workshop.

Housen, A. & F. Kuiken (2009). Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics* 30(4), 461–473. URL http://applij.oxfordjournals.org/content/30/4/461.full.pdf.

Lakshmanan, U. & L. Selinker (2001). Analysing interlanguage: how do we know what learners know? *Second Language Research* 17(4), 393–420. URL http://proxy.lib.ohio-state.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ufh&AN=7393417&site=ehost-live.

Lee, J. & S. Seneff (2006). Automatic Grammar Correction for Second-Language Learners. In *INTERSPEECH 2006 – ICSLP*. URL http://groups.csail.mit.edu/sls/publications/2006/IS061299.pdf.

On emergent linguistic characteristics in learner and translation corpora
Detmar Meurers
Introduction
Background
Learner Corpora
Data in SLA Research
Corpus annotation
Categories for Learner Language
Example: Parts-of-speech
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence
Systematic categories
Nature of interlang. categories
Comparative fallacy
Syntactic annotation
Error annotation
Target hypotheses
Activity & learner modeling
Task-specific learner corpora
Emerging Units in Translation Corpora
Variation detection
Issues in working out the idea
Conclusion
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
39 / 39

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4), 474–496. URL http://www.ingentaconnect.com/content/jbp/ijcl/2010/00000015/00000004/art00002.

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In P. Grommes & M. Walter (eds.), *Fortgeschrittene Lernervarietäten*, Tübingen: Niemeyer, pp. 119–140.

Lüdeling, A., M. Walter, E. Kroymann & P. Adolphs (2005). Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics*. Birmingham. URL http://www.corpus.bham.ac.uk/PCLC/Falko-CL2006.doc.

MacWhinney, B. (2008). Enriching CHILDES for morphosyntactic analysis. In H. Behrens (ed.), *Corpora in Language Acquisition Research: History, Methods, Perspectives*, Amsterdam and Philadelphia: John Benjamins, vol. 6 of *Trends in Language Acquisition Research*, pp. 165–197. URL http://childes.psy.cmu.edu/grasp/morphosyntax.doc.

Mair, C. & M. Hundt (eds.) (2000). *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi.

Metcalf, V. & D. Meurers (2006a). Generating Web-based English Preposition Exercises from Real-World Texts. URL http://purl.org/net/icall/handouts/eurocall06-metcalf-meurers.pdf. EUROCALL 2006. Granada, Spain. September 4–7, 2006.

Metcalf, V. & D. Meurers (2006b). When to Use Deep Processing and When Not To – The Example of Word Order Errors. URL http://purl.org/net/icall/handouts/calico06-metcalf-meurers.pdf. Pre-conference Workshop on NLP in CALL – Computational and Linguistic Challenges. CALICO 2006. May 17, 2006. University of Hawaii.

Meunier, F. (1998). Computer Tools for Interlanguage Analysis: A Critical Approach. In G. Sylviane (ed.), *Learner English on Computer*, London and New York: Addison Wesley Longman, pp. 19–37.

Meurers, D. (2009). On the Automatic Analysis of Learner Language: Introduction to the Special Issue. *CALICO Journal* 26(3), 469–473. URL http://purl.org/dm/papers/meurers-09.html.

Meurers, D., N. Ott & R. Ziai (2010a). Compiling a Task-Based Corpus for the Analysis of Learner Language in Context. In *Proceedings of Linguistic Evidence*. Tübingen, pp. 214–217. URL http://purl.org/dm/papers/meurers-ott-ziai-10.html.

Meurers, D., R. Ziai, L. Amaral, A. Boyd, A. Dimitrov, V. Metcalf & N. Ott (2010b). Enhancing Authentic Web Pages for Language Learners. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-5) at NAACL-HLT 2010*. Los Angeles: Association for Computational Linguistics. URL http://purl.org/dm/papers/meurers-ziai-et-al-10.html.

Meurers, W. D. (2005). On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. *Lingua* 115(11), 1619–1639. URL http://purl.org/dm/papers/meurers-03.html.

Meurers, W. D. & S. Müller (2009). Corpora and Syntax (Article 42). In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics*, Berlin: Mouton de Gruyter, vol. 2 of *Handbooks of Linguistics and Communication Science*, pp. 920–933. URL http://purl.org/dm/papers/meurers-mueller-09.html.

Miura, S. (1998). Hiroshima English Learners' Corpus: English learner No. 2 (English I & English II). URL http://home.hiroshima-u.ac.jp/d052121/eigo2.html. Last Modified 14 May, 1998.

On emergent linguistic characteristics in learner and translation corpora
Detmar Meurers
Introduction
Background
Learner Corpora
Data in SLA Research
Corpus annotation
Categories for Learner Language
Example: Parts-of-speech
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence
Systematic categories
Nature of interlang. categories
Comparative fallacy
Syntactic annotation
Error annotation
Target hypotheses
Activity & learner modeling
Task-specific learner corpora
Emerging Units in Translation Corpora
Variation detection
Issues in working out the idea
Conclusion
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
39 / 39

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* 24(4), 492—-518.

Ott, N. (2009). Information Retrieval for Language Learning: An Exploration of Text Difficulty Measures. ISCL master's thesis, Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen, Germany. URL http://drni.de/zap/ma-thesis.

Ott, N. & D. Meurers (2010). Information Retrieval for Education: Making Search Engines Language Aware. *Themes in Science and Technology Education. Special issue on computer-aided language analysis, teaching and learning: Approaches, perspectives and applications* 3(1–2), 9–30. URL http://purl.org/dm/papers/ott-meurers-10.html.

Ott, N. & R. Ziai (2010). Evaluating Dependency Parsing Performance on German Learner Language. In M. Dickinson, K. Müürisep & M. Passarotti (eds.), *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*. vol. 9 of *NEALT Proceeding Series*, pp. 175–186. URL http://www.sfs.uni-tuebingen.de/~rziai/papers/Ott.Ziai-10.pdf.

Rosén, V. & K. D. Smedt (2010). Syntactic Annotation of Learner Corpora. In H. Johansen, A. Golden, J. E. Hagen & A.-K. Helland (eds.), *Systematisk, variert, men ikke tilfeldig. Antologi om norsk som andrespråk i anledning Kari Tenfjords 60-årsdag [Systematic, varied, but not arbitrary. Anthology about Norwegian as a second language on the occasion of Kari Tenfjord's 60th birthday]*, Oslo: Novus forlag, pp. 120–132.

Sampson, G. & A. Babarczy (2003). Limits to annotation precision. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*. pp. 61–68. URL http://www.sfs.uni-tuebingen.de/~zinsmeis/AnnotCorp05/materials/sampson-barbarczy03.pdf.

On emergent linguistic characteristics in learner and translation corpora
Detmar Meurers
Introduction
Background
Learner Corpora
Data in SLA Research
Corpus annotation
Categories for Learner Language
Example: Parts-of-speech
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence
Systematic categories
Nature of interlang. categories
Comparative fallacy
Syntactic annotation
Error annotation
Target hypotheses
Activity & learner modeling
Task-specific learner corpora
Emerging Units in Translation Corpora
Variation detection
Issues in working out the idea
Conclusion
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
39 / 39

Tenfjord, K., J. E. Hagen & H. Johansen (2006). The Hows and Whys of coding categories in a learner corpus (or "How and Why an error-tagged learner corpus is not ipso facto one big comparative fallacy"). *Rivista di psicolinguistica applicata* 6, 93–108.

Tetreault, J. & M. Chodorow (2008). The Ups and Downs of Preposition Error Detection in ESL Writing. In *Proceedings of COLING-08*. Manchester, UK. URL http://www.ets.org/Media/Research/pdf/r3.pdf.

van Rooy, B. & L. Schäfer (2002). The Effect of Learner Errors on POS Tag Errors during Automatic POS Tagging. *Southern African Linguistics and Applied Language Studies* 20, 325–335.

van Rooy, B. & L. Schäfer (2003). An Evaluation of Three POS Taggers for the Tagging of the Tswana Learner English Corpus. In D. Archer, P. Rayson, A. Wilson & T. McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 conference Lancaster University (UK), 28 – 31 March 2003*. vol. 16 of *University Centre For Computer Corpus Research On Language Technical Papers*, pp. 835–844. URL http://www.corpus4u.org/upload/forum/2005092023174960.pdf.

Volk, M., A. Göhring, T. Marek & Y. Samuelsson (2010). SMULTRON (version 3.0) — The Stockholm MULtilingual parallel TReebank. URL http://www.cl.uzh.ch/research/paralleltreebanks_en.html. An English-French-German-Spanish-Swedish parallel treebank with sub-sentential alignments.

Voutilainen, A. & T. Järvinen (1995). Specifying a shallow grammatical representation for parsing purposes. In *Proceedings of the 7th Conference of the EACL*. Dublin, Ireland. URL http://portal.acm.org/ft_gateway.cfm?id=977003&type=pdf&coll=GUIDE&dl=GUIDE&CFID=47108142&CFTOKEN=71182750.

On emergent linguistic characteristics in learner and translation corpora
Detmar Meurers
Introduction
Background
Learner Corpora
Data in SLA Research
Corpus annotation
Categories for Learner Language
Example: Parts-of-speech
Automatic POS-Tagging
Three Sources of Evidence
Mismatching Evidence
Systematic categories
Nature of interlang. categories
Comparative fallacy
Syntactic annotation
Error annotation
Target hypotheses
Activity & learner modeling
Task-specific learner corpora
Emerging Units in Translation Corpora
Variation detection
Issues in working out the idea
Conclusion
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
39 / 39

Wolfe-Quintero, K., S. Inagaki & H.-Y. Kim (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Honolulu: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.

Zyzik, E. & C. Azevedo (2009). Word Class Distinctions in Second Language Acquisition. *SSLA* 31(31), 1–29. URL http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=3981776.