Multilingual and Parallel Corpora
Non-native corpora

Amir Zeldes

amir.zeldes@georgetown.edu

Announcement: Monday

- Laura Vilardell will be talking about working as a translator and Translation Memories:
  - Please install OmegaT, an assisted translation program.
    - To download OmegaT, use the 'download selector', 'traditional', then the operating system of your computer, then 'standard' and then it will be downloaded.
    - Maybe it will alert you that it is an internet site which is not safe, then if you have a Mac you have to go to system preferences and in security and privacy accept the certificate.
- Link: http://www.omegat.org/en/downloads.html
  - E-Mail lv210@georgetown.edu if you have any trouble

Extracting parallel elements - association

- MI3 (Daille 1995):
  Given *a*, how likely is *b*?

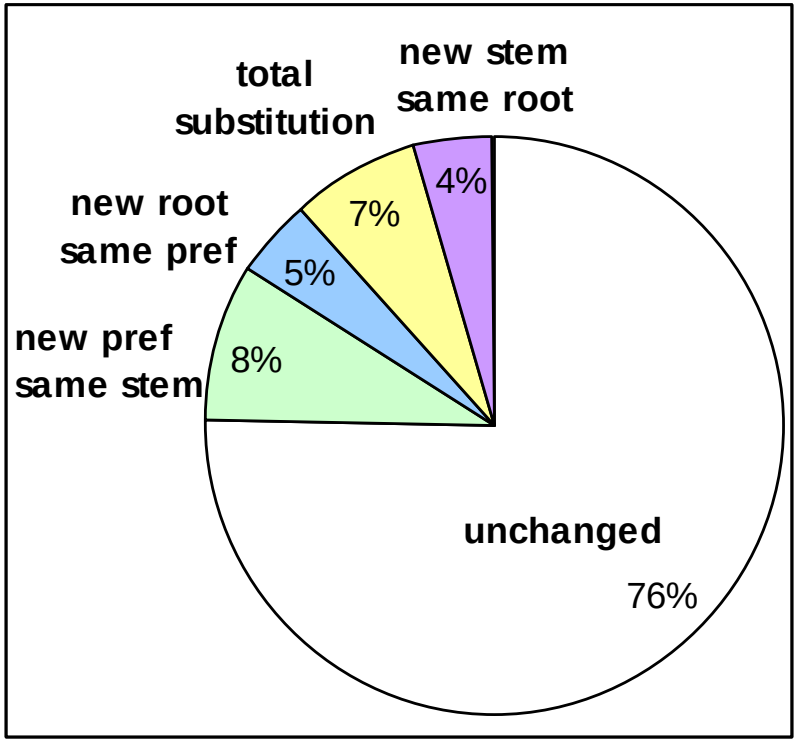
$$MI3 = \log\left(\frac{pairs(a\&b)^3 \cdot N}{pairs(a) \cdot pairs(b)}\right)$$

- Supports **negative** association (beyond MT models)
- Allow *n* to *m* pairs – here, bi-grams to unigrams

a	b	translation	pairs(a)	pairs(b)	pairs(a&b)	MI3
przedni kapłan	arcykapłan	high priest	441	237	19	13.931
słowo	słowo	word	343	585	24	14.001
…						

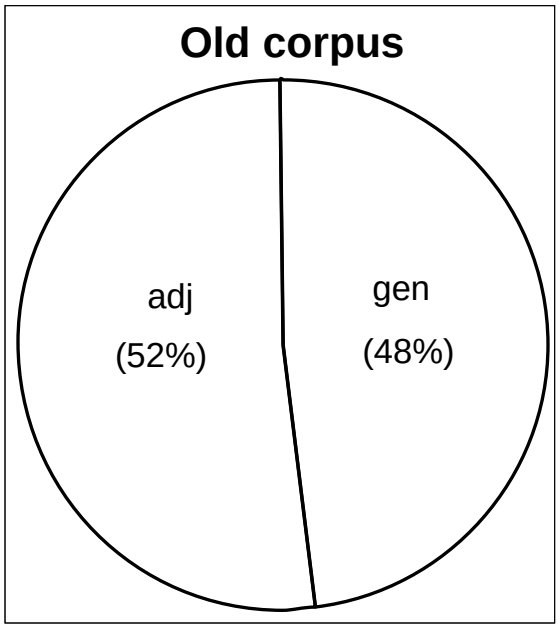
Verb-Verb pairs

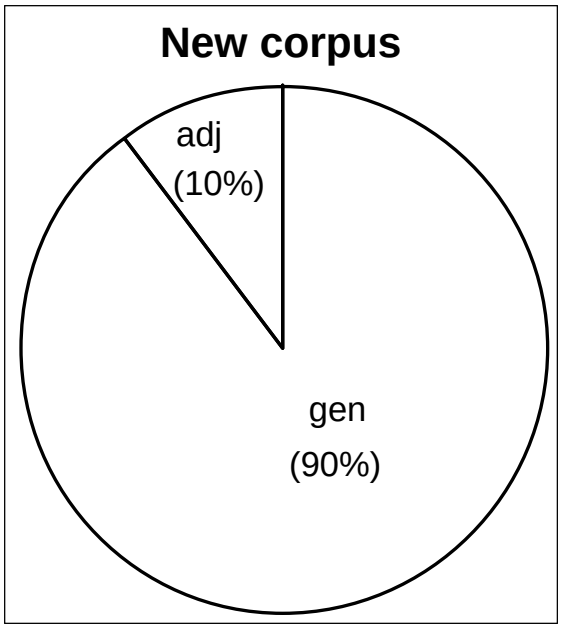
- 76% identical
- 7% replaced
- 8% prefix change
- 5% root change
- 4% stem alteration



Possessive adjective

- Non-aligned distribution is **misleading**!





Non-native corpora

- Corpora of non-native language use are highly valuable:
  - Data on variation within a broad frame ‘English’ etc.
  - Pedagogical applications –
    - Detect, analyze and correct errors
    - Improve teaching materials
    - Use in the classroom
    - Task based teaching
  - Central issue: **aligning** L2 to L1 equivalents

Task based teaching: *the* vs. 0 article

- **Example from:** http://www.eisu.bham.ac.uk/johnstf/def_art.htm

- In Gwynedd, a bedrock of **the Welsh language**, there are 25 film-making companies.
- We must accept that the salvation of **the French language** involves learning one or more of the languages in neighbouring countries.
- The research also showed increases in the frequency of **bad language** and sex on television.
- Inspectors said behaviour was generally good, but features "such as free use of **colloquial language** and non-attendance at lessons are tolerated much more than in conventional schools".

- 1. proud of their command of _____ English language and engage in quite of lot of patting them
- 2. but it does not mean that _____ everyday language is bad: it is simply the way of things tha
- 3. cluded that cerebral dominance for _____ language is established before the age of five. Dur
- 4. abulary is one thing and _____ technical language is another, Vocabulary is words, lists of
- 5. avic-speakers. Orthodoxy and _____ Greek language remain the two markers of modern Greek ide

Data for studying SLA

- Intuition
- Collections of L2 data
  - Unsystematic, anecdotal data
  - Error collections
  - Learner corpora
- Experimental data
  - Elicitation data
  - Psycholinguistic experiments
  - …

Data for studying SLA

- Intuition: of learners? teachers? reliability?
- Collections of L2 data:
  - Unsystematic, anecdotal data – problematic
  - Error collections – problematic
  - Learner corpora
- Experimental data
  - Elicitation data
  - Psycholinguistic experiments
  - …

Learner corpora

- "Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance."
  (Granger 2002: 7)

Annotation challenges

- Collecting unannotated data is easiest
- Annotation with automatic tools (POS tagging)
  - Difficult and problematic
  - Often requires manual correction
- Ideally: Error annotation

Data types

- Essays (most common, best for more advanced students)
- Summaries – often problematic
- Task based – potentially problematic
- Longitudinal

- Key issues:
  - Metadata collection
  - Proficiency estimation – ideally independent from text

Task prompt text reuse

- *Sprecher nehmen an, daß Unterschiede hinsichtlich der Form auch Unterschiede in der Wortbedeutung signalisieren.*
- Speakers assume that differences in form also signal differences in word meaning

- *Sprecher nehmen an, dass Unterschiede hinsichtlich der Form auch Unterschiede in der Wortbedeutung signalisieren.*
- *Kontrast bedeutet also: Sprecher nehmen an, dass Unterschiede hinsichtlich der Form auch Unterschiede in der Wortbedeutung signalisieren.*

Collection methods

Hand written

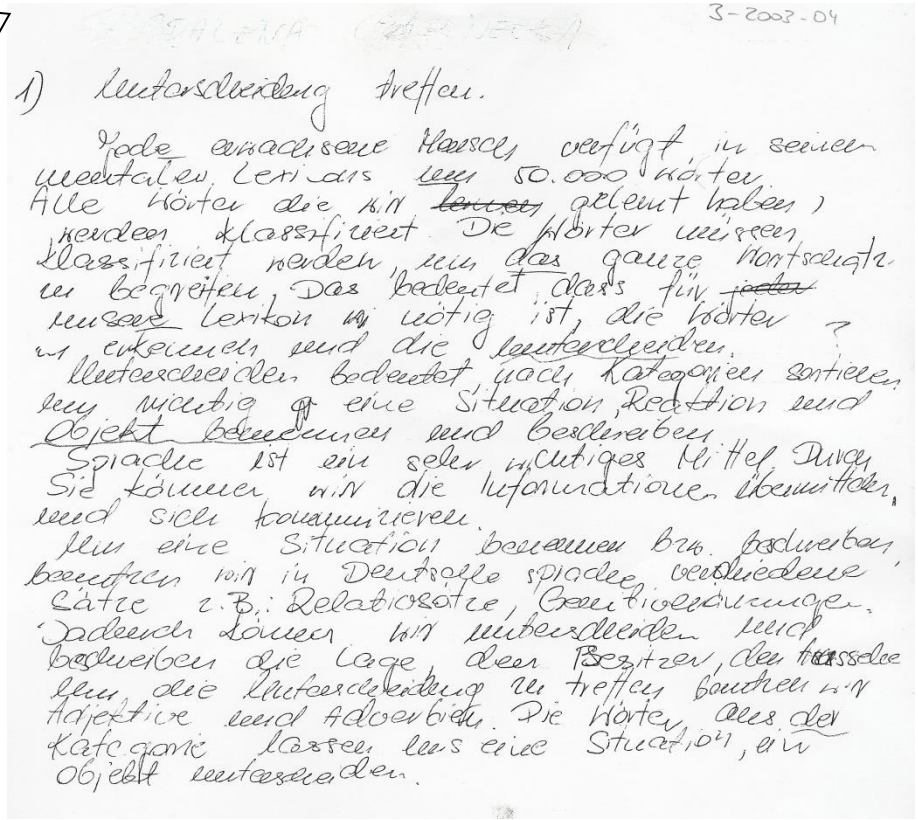
- Rules out typos
- No chance of spellchecking, uncontrollable resources
- More difficult to digitize, potentially problematic

Digital

- Computer medium errors possible (typos, typographical issues)
- Difficult to rule out use of software, find/replace…
- Easy to process

Challenges

anonymization!



3-2003-04

1) Wortunterscheidung treffen.

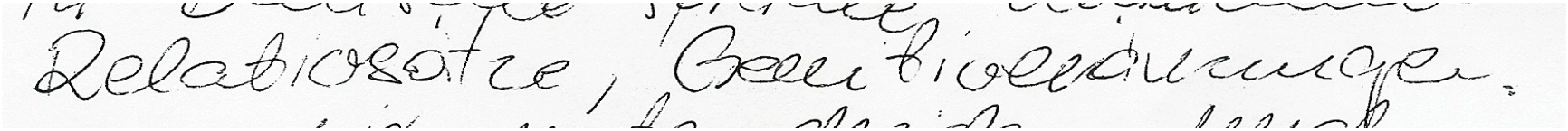
Jede erwachsene Person verfügt in seiner Muttersprache über ca. 50.000 Wörter. Alle Wörter die wir kennen gelernt haben, werden klassifiziert. Die Wörter müssen klassifiziert werden, um das ganze Wortschatz zu begreifen. Das bedeutet dass für jeden unser Lexikon es nötig ist, die Wörter zu erkennen sind die Wortunterscheiden. Wortunterscheiden bedeutet grob Kategorien sortieren um wichtig zu eine Situation, Reaktion und Objekt bezeichnen und beschreiben. Sprache ist ein sehr wichtiges Mittel. Durch Sie können wir die Informationen übermitteln, und sich kommunizieren.

Um eine Situation benennen bzw. beschreiben wir in Deutsche Sprache verschiedene Sätze z.B.: Relativsätze, Genitivkonstruktionen. Dadurch können wir unterscheiden und beschreiben die Lage, den Besitzer, den Grund. Um die Wortunterscheidung zu treffen benutzen wir Adjektive und Adverbien. Die Wörter aus der Kategorie lassen uns eine Situation, ein Objekt unterscheiden.

metadata

Analysis challenges

How to avoid subconscious correction?



- Digitization implies research-relevant decisions!

Error annotation – goals

- Examine errors/deviations in L2 grammar
- One options is to **tag error types**
  - Enables easy search
  - Standardization of error taxonomy
  - Quantitative evaluation

What is an error?

- Traditional view – violation of a rule
  - Ungrammatical
  - Where do we get the rules from?
    - Explicitly formulable
    - Implicitly noticeable
- Deviation from a norm, "breaches of code" (Corder 1973)
  - Unacceptable/inappropriate
  - Multiple norms possible
  - Not necessarily decidable

What is an error?

- "A linguistic form, ... which, in the same context would in all likelihood not be produced by the learner's native speaker counterparts." (Lennon 1991, 182)

- Ring a bell?
- Is this a usable definition?
- Need for reproducibility

What is an error?

- Sometimes a difference is postulated between *error* (≈ Competence Error) and *mistake/lapse* (≈ Performance Error)
  - How can we recognize these in a corpus?
  - Qualitative distinction?

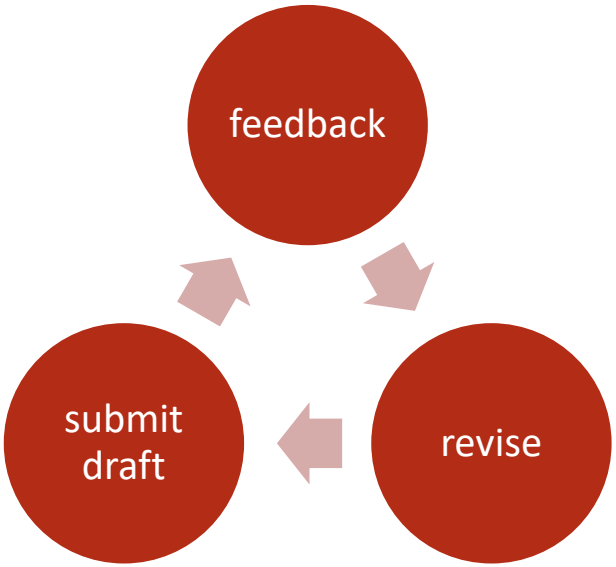
Quick exercise

- Find the errors in the following L2 sentence: (adapted from Lüdeling 2008)
  - *It's a type of problem, what the novel or ode not applying.*

- What kind of error types can we postulate?
- Is there a list of possible error types?

Case study: aligned error annotation

- Hong Kong City University Corpus of English Learner Academic Drafts (Lee et al. 2015)
  - 7.7 M tokens
  - Approx. 11 K essay **drafts** (about 4,300 essays)
  - L1 mostly Cantonese
- Concept:



Data collection

Clause Level

Comment	**Relative pronoun - missing**
Explanation	You need to link these phrases with relative pronouns
Examples of Wrong Use	The student gave the presentation made some interesting points.
Correct Use	The student *who* gave the presentation made some interesting points.
External Links	http://owl.english.purdue.edu/owl/resource/645/01/

~~d. They include adding wrong amount of solution in~~
~~er to have a more accurate, the technique needed to be~~

erial cell to burst is 85 minutes. At t=105, this is a dramatic increase of phage titres, reaching to the value of 270million p.f.u./ml. This shows the phenomenon of one-step growth for the T4 bacteriophage.

In each sample time, the phage titres represent the total number of phage that is inside the growth tube, including both the infected virus and [01]suspended virus. Because of it [Be more explicit; what do you refer by 'it'], a sample tube [38]has added in CHCL is used as a control. CHCL can killed infected virus, only the suspensions virus will be counted in the overlay method. By this information, the percentage of viral adsorption can be determined.

In this experiment, error ~~is~~ exist [17]as there is several practical mistake[plural] have been noticed. They include adding wrong amount of solution in the dilution process and using the wrong auto pipette in the mixing procedure. In order to have a more accurate, the technique needed [present tense]to be improved.

[Nice report! You have explained your experiment and your results succesfully. However, it is very important for a Lab report to include background information of previous studies in the Introduction stage, as well as to relate these studies findings with your own results in the Discussion stage. In order to validate your results, to need to support them by referencing other results found in some other studies. In addition, it is essential that in the Discussion stage you make reference to what you stated in your Introduction, so as to reject or support your research purpose. Please bear in mind these comments to prepare assignment 1. I'm looking forward to reading your next piece of work. Best,]

Data collection

Discipline	# Essays	Discipline	# Essays
Applied physics	288	Electronic engineering	460
Asian and international studies	118	General education	172
Biology	618	Law	20
Building science and technology business	249	Linguistics	644
Business	690	Management sciences	414
Computer science	148	Social studies	477
Creative media	39		

Error annotation scheme

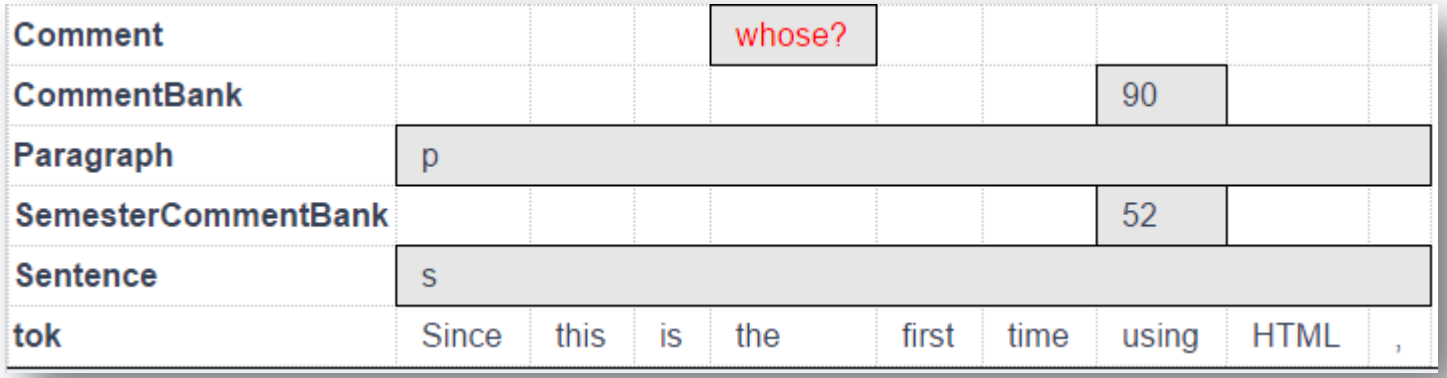
- Closed comment bank (95 categories)

ID	Category
1	Adjective Compar./Superl. Form
47_1	Adjective needed – POS Incorrect
4	Article – unnecessary
5	Article - wrong use
6	Article missing
8	Coherence - Introductory Paragraph Missing
12	Coherence - missing conclusion
19	Coherence - Too many focuses in one paragraph

ID	Category
29	Delete this (unnecessary)
30	Heading - inappropriate
31	Heading - missing
40	Modal - missing
41	Modal - wrong use
42	Noun - countable
61	Question - Do support
72	Subject - Dummy subject

Error annotation scheme

- Open ended comments
  - spacing
  - use singular form
  - as compared to when?
  - we do not use MUCH alone, usually we say NOT MUCH, SO MUCH



Which errors are frequent?

	# Comments
Essay-level error categories	
Informal language	1321
Coherence—more elaboration is needed	655
Paragraph—new paragraph	516
Coherence—signposting	315
Coherence—missing topic sentence	191
Clause-level error categories	
Punctuation—missing	2371
Conjunction—missing	1874
Word order	1577
Punctuation—capitalisation	1475
Sentence—new sentence	1345
Word-level error categories	
Article—missing	10,280
Delete this (unnecessary)	9109
Noun—countable	7066
Subject–verb agreement	3929
Preposition—wrong use	3718

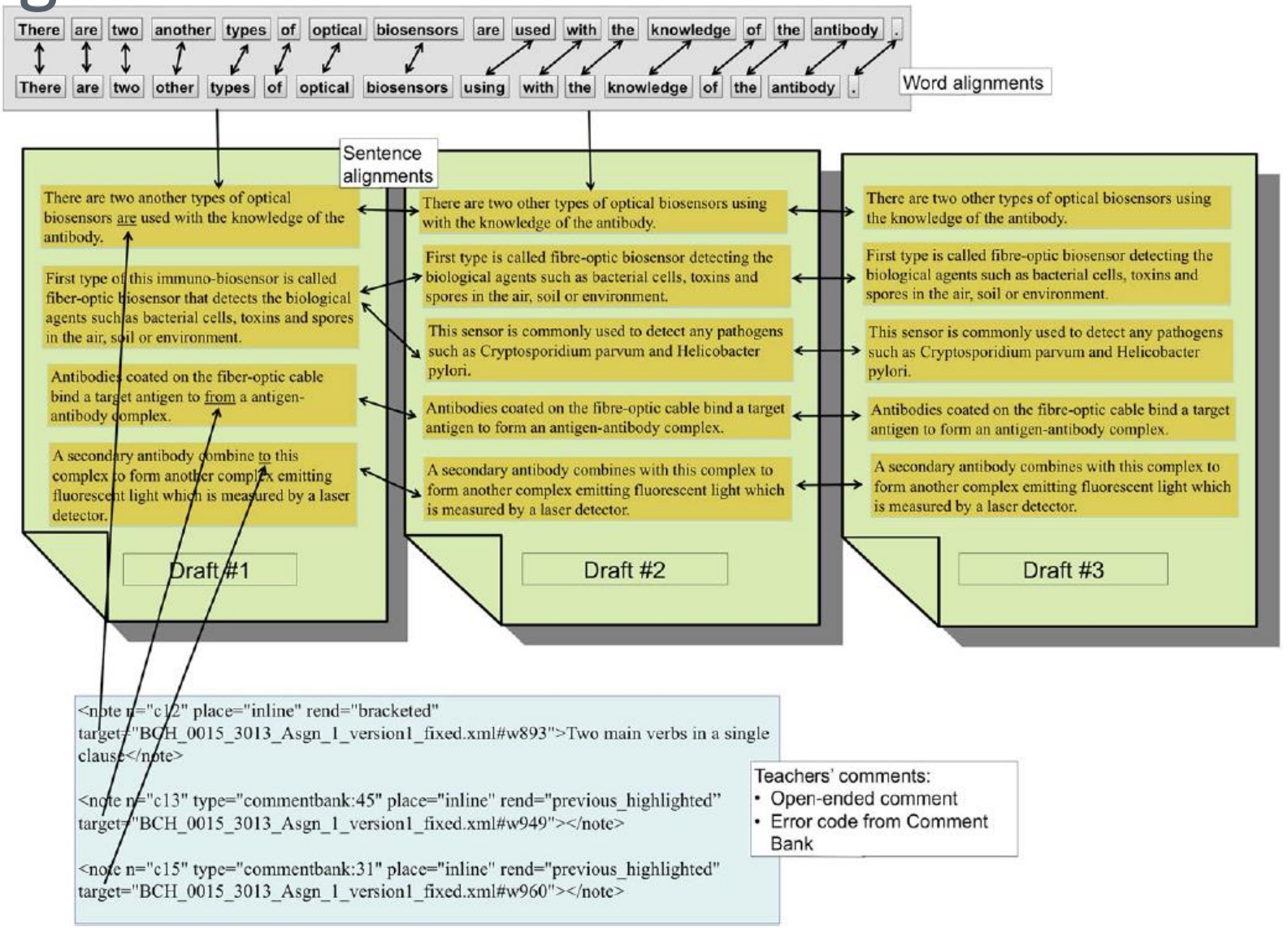
Are error categories decidable?

- Experiment:
  - 200 sentences
  - 2 annotators

Error category	Agreement level (%)
Article—missing	87.1
Noun—countable	78.4
Delete this	78.9
Subject–verb agreement	78.8
Preposition—wrong use	73.9

- How can we interpret this?
- What can we do about it?

Alignment


There are two another types of optical biosensors are used with the knowledge of the antibody .
There are two other types of optical biosensors using with the knowledge of the antibody .
Word alignments
Sentence alignments
There are two another types of optical biosensors are used with the knowledge of the antibody.
First type of this immuno-biosensor is called fiber-optic biosensor that detects the biological agents such as bacterial cells, toxins and spores in the air, soil or environment.
Antibodies coated on the fiber-optic cable bind a target antigen to from a antigen-antibody complex.
A secondary antibody combine to this complex to form another complex emitting fluorescent light which is measured by a laser detector.
There are two other types of optical biosensors using with the knowledge of the antibody.
First type is called fibre-optic biosensor detecting the biological agents such as bacterial cells, toxins and spores in the air, soil or environment.
This sensor is commonly used to detect any pathogens such as Cryptosporidium parvum and Helicobacter pylori.
Antibodies coated on the fibre-optic cable bind a target antigen to form an antigen-antibody complex.
A secondary antibody combines with this complex to form another complex emitting fluorescent light which is measured by a laser detector.
There are two other types of optical biosensors using the knowledge of the antibody.
First type is called fibre-optic biosensor detecting the biological agents such as bacterial cells, toxins and spores in the air, soil or environment.
This sensor is commonly used to detect any pathogens such as Cryptosporidium parvum and Helicobacter pylori.
Antibodies coated on the fibre-optic cable bind a target antigen to form an antigen-antibody complex.
A secondary antibody combines with this complex to form another complex emitting fluorescent light which is measured by a laser detector.
Draft #1
Draft #2
Draft #3
<note n="c12" place="inline" rend="bracketed" target="BCH_0015_3013_Asgn_1_version1_fixed.xml#w893">Two main verbs in a single clause</note>
<note n="c13" type="commentbank:45" place="inline" rend="previous_highlighted" target="BCH_0015_3013_Asgn_1_version1_fixed.xml#w949"></note>
<note n="c15" type="commentbank:31" place="inline" rend="previous_highlighted" target="BCH_0015_3013_Asgn_1_version1_fixed.xml#w960"></note>
Teachers' comments:
• Open-ended comment
• Error code from Comment Bank

How to align sentences?

Suppose two sentences, at similar positions in both drafts, share a considerable number of words. The first sentence might have been edited into the second, in which case they should be aligned; alternatively, the first sentence might have been simply deleted and the second inserted, in which case they should not be aligned.

Our principle was to prefer ***higher recall*** *of alignments at the risk of lower precision—i.e. to align sentence pairs with relatively low similarity—since it is much easier for the corpus user to discount an alignment than to recover an unidentified alignment. This policy was enforced by setting* ***a relatively high cost for insertion and deletion, merge and split****.*

Accuracy:
89.8 %

Word alignment

- Use Translation Error Rate (Snover et al. 2006), allowing edits: insert, delete, replace, shift
  - Annotate words as ‘to be deleted’ if they will disappear in next version
  - Annotate as ‘inserted’ if they were introduced in this version

Thus	the	user	can	be	quickly	to	acquire	and	familiarize	(	pronoun	missing
				del	del	del		del	del	del	del	del
ins	ins	ins	ins	ins	ins	ins	ins	ins	ins	ins	ins	ins
thus	the	user	can	be	quickly	to	acquire	and	familiarize	(	pronoun	miss
RB	DT	NN	MD	VB	RB	TO	VB	CC	VB	CD	NN	VBG

.	Thus	the	user	can	acquire	the	essential	operation	of	HTML
del	del	del	del	del	del	del	del	del	del	del
.	thus	the	user	can	acquire	the	essential	operation	of	HTML
.	RB	DT	NN	MD	VB	DT	JJ	NN	IN	NNP

Edge annotations

- Just like annotating words, we can annotate the links (edges) connecting sentences and words across versions:

Attribute	Sentence alignment	Word alignment
From draft	The version number of the draft from which the sentences are aligned.	The version number of the draft the words are aligned from
To draft	The version number of the draft to which the sentences are aligned.	The version number of the draft the words are aligned to
Type	'Identical', 'replace', 'merge' or 'split'	'Identical', 'replace' or 'shift'

- Queriable at:
  - https://corpling.uis.georgetown.edu/annis/

Do errors decrease?

- In each aligned version, the text is supposed to get 'better'
- Not all comments signal errors, but:

Table 4 The number of comments in various stages of the revision cycle

Comment type	Draft #1	Draft #2	Draft #3+
Open-ended comments (per 100 words)	33,534 (1.24)	17,597 (0.87)	1304 (0.15)
Error categories (per 100 words)	60,875 (2.26)	24,341 (1.20)	1379 (0.16)

- However: **unaligned** data
- Do comments actually improve drafts?

Case study: tense revision

- Does feedback actually improve specific problems?
- NB: **aligned** data

→ Suggested tense ↓ Tense in draft	Present simple	Past simple	Present perfect	Past perfect
Present simple	–	954	89	46
Past simple	345	–	92	21
Past/present perfect	133	131	–	–
Continuous	198	60	12	8
Base form	96	222	49	26
Total	772	1367	242	101
% of changes	76.94 %	74.91 %	71.07 %	42.57 %
% of correct changes	56.09 %	60.35 %	42.56 %	20.79 %

Reliability and accountability

- Error annotation can be very useful:
  - Direct access to linguistic categories of interest
    - Articles, tenses, clauses…
  - Enable feedback to learners
- But there is a reliability problem:
  - Agreement is only in the 80s
  - Especially problematic for some of the more interesting errors
  - No accountability – why did one annotator choose error type A and another error B?

Target hypotheses

- It's not really possible to mark errors without having a **target hypothesis** in your head

- Errors are recognized by comparison to your hypothesis
  - Implicit or explicit?
  - Agreement?

Target hypotheses

- Reading:
  - Reznicek et al. (2013) Competing Target Hypotheses in the Falko Corpus: A Flexible Multi-Layer Corpus Architecture