Linguistic annotations

Hakyung Sung
University of Hawaii at Manoa

October 21, 2026

Topics to cover

- Review (10 mins)
- Presentation (20 mins)
- Linguistic annotations (30 mins)
  - What
  - Why
  - How
- Lab: Annotation exercise

Review:
Measuring constructs in SLA

Learner corpus research - SLA

Goal

To turn learner language into empirical evidence about SLA.

- Learner corpora contain *observable* language production.
- SLA research asks questions about *theoretical* constructs.

Learner corpus research - SLA

Complexity, accuracy, fluency are well-known theoretical constructs in SLA.

To investigate them in corpora, we need to decide:

- What exactly does the construct mean?
  operationalization

- Which linguistic features represent it?
  e.g., syntactic hierarchies (words in clauses), grammatical categories (noun - noun modifiers)

- How will those features be measured?
  e.g., length (mean length of X), diversity (different types of grammatical structures in X)

Development involves more than accuracy

The complexity–accuracy–fluency (CAF) framework distinguishes several aspects of learner performance (Norris & Oretega, 2009).

Accuracy	Conformity to target-language norms.
Fluency	Efficiency and automatization of language use.
Complexity	**Expansion and elaboration of linguistic resources.**

Complexity

- Syntactic complexity (some examples in last week's presentation)
- Grammatical complexity (today's presentation)

Complexity: Examples

- A: The students very carefully reviewed the report yesterday afternoon.
- B: The students reviewed the report on local education policy.
- Both: 9 words, 1 clause $\rightarrow$?
- Grammatical structures
  - A: ?
  - B: ?
- Syntactic and grammatical complexity measures may reveal different aspects of learner production (with patterns varying across *spoken* and *written* modes and across *registers*).

Presentation: Biber & Larsson (2026)

Linguistic annotations: What

What is linguistic annotation

- Corpus analysis often involves more than raw text.
- We add linguistic information to identify and analyze patterns.
- Linguistic annotation = labeling language data with information.

Fruit flies like an apple.

Fruit flies like an apple.

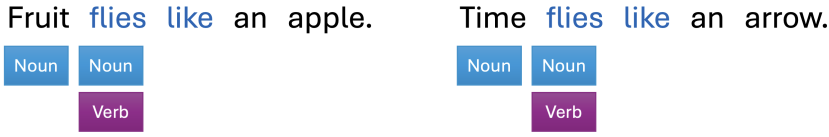
Time flies like an arrow.

Fruit flies like an apple.

Noun

Time flies like an arrow.

Noun


Fruit flies like an apple.
Noun
Noun
Verb
Time flies like an arrow.
Noun
Noun
Verb

Fruit flies like an apple.



Time flies like an arrow.




Fruit flies like an apple.
Noun
Noun
Prep
Det
Noun
Verb
Verb


Time flies like an arrow.
Noun
Noun
Prep
Det
Noun
Verb
Verb





Make a decision based on linguistic interpretation of the given dataset

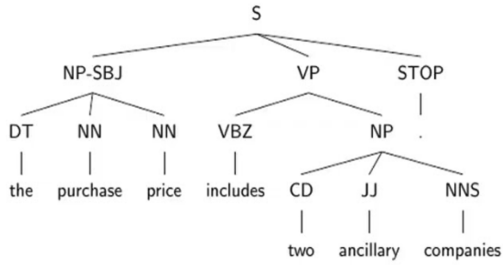
Different types of linguistic annotations

- **Type: Syntactic information**

- *English Penn Treebank corpus*
  - Used for character-level and word-level *language modeling*

Table 2
The Penn Treebank POS tagset.

1.	CC	Coordinating conjunction	
2.	CD	Cardinal number	
3.	DT	Determiner	
4.	EX	Existential *there*	
5.	FW	Foreign word	
6.	IN	Preposition/subordinating conjunction	
7.	JJ	Adjective	
8.	JJR	Adjective, comparative	
9.	JJS	Adjective, superlative	
10.	LS	List item marker	
11.	MD	Modal	
12.	NN	Noun, singular or mass	
13.	NNS	Noun, plural	
14.	NNP	Proper noun, singular	
15.	NNPS	Proper noun, plural	
16.	PDT	Predeterminer	
17.	POS	Possessive ending	
18.	PRP	Personal pronoun	
19.	PP$	Possessive pronoun	
20.	RB	Adverb	
21.	RBR	Adverb, comparative	
22.	RBS	Adverb, superlative	
23.	RP	Particle	
24.	SYM	Symbol (mathematical or scientific)	
25.	TO	*to*	
26.	UH	Interjection	
27.	VB	Verb, base form	
28.	VBD	Verb, past tense	
29.	VBG	Verb, gerund/present participle	
30.	VBN	Verb, past participle	
31.	VBP	Verb, non-3rd ps. sing. present	
32.	VBZ	Verb, 3rd ps. sing. present	
33.	WDT	*wh*-determiner	
34.	WP	*wh*-pronoun	
35.	WP$	Possessive *wh*-pronoun	
36.	WRB	*wh*-adverb	
37.	#	Pound sign	
38.	$	Dollar sign	
39.	.	Sentence-final punctuation	
40.	,	Comma	
41.	:	Colon, semi-colon	
42.	(	Left bracket character	
43.	)	Right bracket character	
44.	"	Straight double quote	
45.	‘	Left open single quote	
46.	“	Left open double quote	
47.	’	Right close single quote	
48.	”	Right close double quote	



Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

More examples of syntactic annotation resources

Resource	Annotation
UD Treebank	Shared POS tags, morphological features, and dependency relations
Prague Treebank	Originally annotated for Czech using the Prague framework (project begun in 1996); available in converted UD form
KAIST Treebank	Originally annotated with constituency structures for Korean (described in Choi et al., 1994); available in converted UD form

- Hajič, J., & Hajičová, E. (1997). Syntactic tagging in the Prague Treebank.
- Choi, K.-S., Han, Y. S., Han, Y. G., & Kwon, O. W. (1994). KAIST Tree Bank Project for Korean: Present and Future Development. In *Proceedings of the International Workshop on Sharable Natural Language Resources* (pp. 7–14).

More and more examples (of different annotations)

Table 9: Representative types of linguistic annotation.

Type of annotation	Description	Examples
POS tagging	Adding information about the grammatical category (part of speech) of each word in a text.	Noun, verb, adjective, adverb.
Phonetic annotation	Adding information about how words or sounds are pronounced in a spoken corpus.	IPA transcription, individual speech sounds.
Prosodic annotation	Adding information about the prosodic features of spoken language.	Stress, intonation, pauses, rhythm.
Syntactic annotation	Adding information about the syntactic structure of sentences, including how words are organized into phrases and clauses.	Constituency parsing, dependency parsing.
Semantic annotation	Adding information about the meanings or semantic categories of words and expressions.	*Cricket* as a sport vs. *cricket* as an insect.

Leech (2004). Adding Linguistic Annotation. Section 2

More and more examples (of different annotations)

Discourse annotation	Adding information about relationships between linguistic elements across sentences or larger stretches of text.	Anaphoric links between pronouns and their antecedents.
Stylistic annotation	Adding information about stylistic features, including how speech and thought are represented in a text.	Direct speech, indirect speech, free indirect thought.
Lexical annotation	Adding information about the lexical identity of words, such as the lemma associated with each word form.	*Running*, *runs*, *ran* $\rightarrow$ *run*.
Error annotation	Adding information about errors in language production, particularly in learner corpora, including their locations and types.	Grammatical errors, lexical errors, spelling errors, word-order errors.

Leech (2004). Adding Linguistic Annotation. Section 2

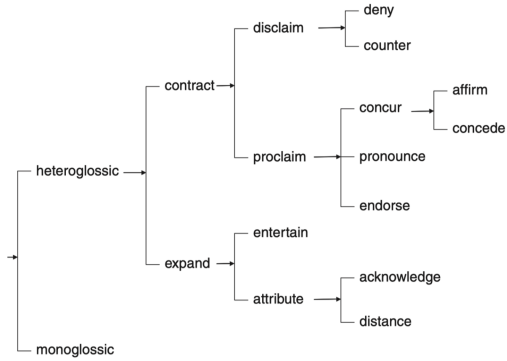
Reinventing the wheel

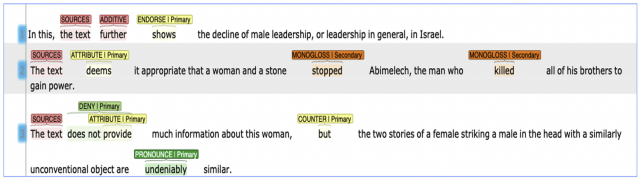
- What if existing annotation schemes do not meet our needs?
- Many existing schemes annotate individual tokens or relations between them:
  - POS tags: Grammatical categories assigned to tokens
  - Dependency structures: Syntactic relations between tokens

MORE examples :)

Target linguistic structures (beyond word-level)

- Stance-taking (Engagement) movement in writing (Eguchi & Kyle, 2023)



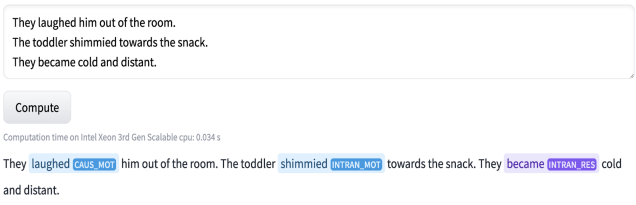


MORE examples :)

Target linguistic structures (beyond word-level)

- Stance-taking (Engagement) movement in writing (Eguchi & Kyle, 2023)
- Argument structure constructions (Kyle & Sung, 2023; Sung & Kyle, 2024)

ASC tag	Syntactic frame (example)	Semantic frame (example)
ATTR	Subj-$V_{copular}$-NP/AdjP/PP	X_{theme} is $Y_{attribute}$
INTRAN_S	Subj-V	X_{agent} acts
INTRAN_MOT	Subj-V-Adverbial Particle/PP	X_{theme} moves Y_{path}
INTRAN_RES	Subj-V-NP/AdjP	$X_{patient}$ becomes Y_{result}
TRAN_S	Subj-V-Obj	X_{agent} acts on $Y_{patient}$
DITRAN	Subj-V-$Obj_{indirect}$-Obj_{direct}	X_{agent} causes $Y_{recipient}$ to receive Z_{theme}
CAUS_MOT	Subj-V-Obj-PP	X_{agent} causes Y_{theme} to move $Z_{path/goal}$
TRAN_RES	Subj-V-Obj-NP/AdjP	X_{agent} causes Y_{theme} to become Z_{state}
PASSIVE	Subj-aux$V_{past\ participle}$ (*-by*-PP)	X_{theme} undergo V (*by* Y_{agent})



MORE examples :)

Target linguistic structures (beyond word-level)

- Stance-taking (Engagement) movement in writing (Eguchi & Kyle, 2023)
- Argument structure constructions (Kyle & Sung, 2023; Sung & Kyle, 2024)
- Lexicogrammatical features (Kyle et al., *in prep*)

idx	word	lemma	cx tag	tag	cat1	cat2	cat3	cat4	cat5	cat6	cat7	cat8	xpos	upos	deprel	headid
0	This	this		pro	dem	sg							DT	PRON	nsubj	1
1	is	be		vbmain	be	pres	simple	active					VBZ	AUX	ROOT	1
2	a	a		dt	art								DT	DET	det	5
3	very	very	rb+jjrbmod	rb	adjmod								RB	ADV	advmod	4
4	important	important	attr+npremod	jj	attr								JJ	ADJ	amod	5
5	opportunity	opportunity		nn		nom							NN	NOUN	attr	1
6	that	that		relpro	relpro_that	sg							WDT	PRON	nsubj	8
7	only	only	rb+advl	rb	advl	ly							RB	ADV	advmod	8
8	comes	come	finitecls+rel	vbmain	vblex	pres	simple	active	nmod_cls	thatcls	rel		VBZ	VERB	relcl	5
9	once	once	rb+advl	rb	advl								RB	ADV	advmod	8
10	in	in	in+advl	in	advl								IN	ADP	prep	8
11	a	a		dt	art								DT	DET	det	12
12	lifetime	lifetime		nn									NN	NOUN	pobj	10
13	.	.											.	PUNCT	punct	1

Linguistic annotations: Why

Benefits of linguistic annotation

Beyond the sheer joy of annotating…

Benefits of linguistic annotation

- Efficient retrieval: Search for linguistic structures beyond word forms
  - They *separate* the groups. → Verb
  - They live in *separate* houses. → Adjective
  - Dependency annotation allows searches for arguments of particular verbs.

Benefits of linguistic annotation

- Reusability: Annotate once, investigate many questions
  - Annotation first requires substantial time and effort
  - Shared resources reduce duplicated effort and support comparisons across studies (benchmarks for language models)

Benefits of linguistic annotation

- cf. data labeling: assigning labels to data, often to train and evaluate machine learning models

- Provide labeled data for language-related machine learning tasks (recall supervised learning); build your own model!

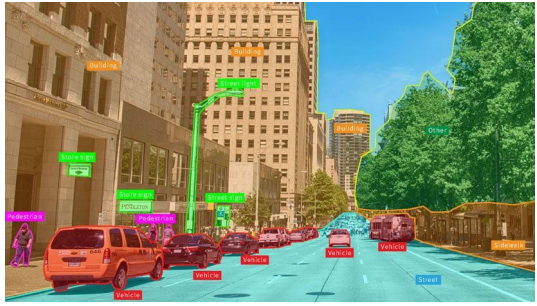


Image source from `https://kr.appen.com/blog/data-annotation/`

Principles of good annotation practice

- Preserve the original data
  - Keep the original text or recording recoverable.
- Document the annotation process
  - Specify the scheme, tools and model versions, annotators, and annotation stages.
  - Record training, quality checks, revisions, and disagreement resolution.

Linguistic annotations: How

Evaluating annotation

- **Validity**: Does the scheme capture meaningful linguistic distinctions?
- **Accuracy**: Do assigned labels match the reference labels?
- **Consistency**: Do independent annotators apply the scheme in the same way?

Validity

- Consistent labeling != a linguistically meaningful annotation scheme.

Validity

- Human interpretation: Annotators may interpret the same example differently.
  - We often recruit undergraduate linguistics majors (and/or crowd sourcing).
  - Even with linguistic training, annotators may disagree.
- Unclear boundaries: Linguistic categories are not always clear-cut.

Example: *Kick the Bucket*

- **Task**: Label constructions as `transitive` or `intransitive`.
- **Case 1**: *kick the ball*
  - Verb + direct object $\rightarrow$ `transitive`
- **Case 2**: *kick the bucket*
  - Literal reading: Physically kick a bucket.
  - Idiomatic reading: Die.
  - Should the interpretation affect our label?
- What are we annotating: **syntactic form**, **meaning**, or both?

Example: *Kick the Bucket*

- **Option 1: Annotate syntactic form**
  - Label both expressions as `transitive`: both have a verb + direct object structure.
- **Option 2: Annotate form and idiomaticity separately**
  - *kick the ball*: `transitive`, `non-idiomatic`
  - *kick the bucket* (meaning "die"): `transitive`, `idiomatic`
  - Idiomaticity and transitivity are different dimensions.
- **Document the decision**: Define the categories, provide examples, and explain difficult cases (from the informed theory/framework)

Accuracy

- Compare assigned labels with a reference or “gold” standard.
- Accuracy: The proportion of evaluated units labeled correctly.

$$A = \frac{C}{N}$$

- C: Number of correctly labeled units
- N: Total number of evaluated units
- Accuracy score can hide poor performance on particular categories.
- Precision and recall help identify different types of error.

Precision/Recall: Identifying nouns

- Suppose `NOUN` is the target category.

	Gold: NOUN	**Gold: Not NOUN**
Predicted: NOUN	True positive (TP)	False positive (FP)
Predicted: Not NOUN	False negative (FN)	True negative (TN)

- **Precision**: Of the tokens labeled as nouns, how many really are nouns?

$$P = \frac{TP}{TP + FP}$$

- **Recall**: Of all actual nouns, how many did the model identify?

$$R = \frac{TP}{TP + FN}$$

Worked Example: Two Noun Taggers

- A corpus contains 100 actual noun tokens.

	Model A	**Model B**
Tokens labeled as nouns	50	150
Correctly identified nouns (TP)	48	95
Incorrect noun predictions (FP)	2	55
Missed nouns (FN)	52	5
Precision	$48/50 = .96$	$95/150 \approx .63$
Recall	$48/100 = .48$	$95/100 = .95$

- Model A: Few false positives, but many missed nouns.
- Model B: Finds most nouns, but includes many non-nouns.
- Which model better serves your research question?

F1 and Performance Across Categories

- F1: The harmonic mean of precision and recall.

$$F_1 = \frac{2PR}{P + R}$$

  - Model A: $F_1 = .64$
  - Model B: $F_1 = .76$

- For multiple categories, report scores by label and specify how overall scores are calculated.
  - *Micro*-F1: Pool counts across categories.
  - *Macro*-F1: Calculate scores for each category, then average them equally.

Consistency

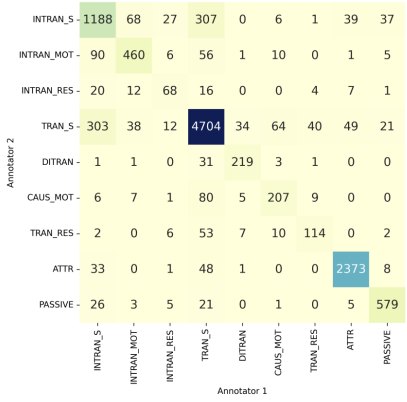
- Do independent human annotators assign the same labels?
- Simple agreement: The proportion of decisions with matching labels.
- Cohen's κ: Adjusts observed agreement for agreement expected from category frequencies.

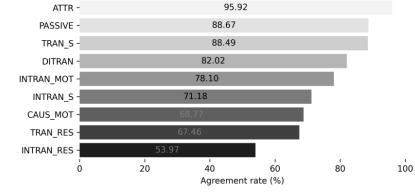
$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

  - P_o: Observed agreement
  - P_e: Expected agreement under the chance model

Consistency

- Use a confusion matrix to identify categories that annotators confuse.





Sung & Kyle (2024). Annotation Scheme for English ASC Treebank.

Lab: Lexicogrammatical feature annotation

Annotation workshop: Goals and tasks

- **Goal**: Annotate four lexicogrammatical features in sample data and develop reusable annotation guidelines.

- **Task 1: Annotate and develop guidelines (20 minutes)**
  - Individually annotate sentences for your assigned feature.
  - Compare annotations within your group and discuss disagreements.
  - Write guidelines with definitions, examples, and rules for difficult cases.

- **Task 2: Exchange and test guidelines (20 minutes)**
  - Read the guidelines from your paired group.
  - Annotate sentences for the new feature only using their guidelines.
  - Compare annotations across groups and revise the guidelines.

- Resources here: `https://hksung.github.io/PythonTutorial/8.html`

Annotation Workshop: Group Assignments

Group	Task 1	Task 2
Group 1	Verb + *that*-complement clause	Noun + *that*-complement clause
Group 2	Noun + *that*-complement clause	Verb + *that*-complement clause
Group 3	*-ing* clause as adverbial	Noun + *-ing* relative clause
Group 4	Noun + *-ing* relative clause	*-ing* clause as adverbial
Group 5	Verb + *-to* complement clause	Adjective + *-to* complement clause
Group 6	Adjective + *-to* complement clause	Verb + *-to* complement clause

- **Paired groups**: Groups 1 and 2; Groups 3 and 4; Groups 5 and 6

Annotation Exercise: Group Debrief

- What worked well, and what challenges did your group encounter?
- Which instructions were helpful? Which were unclear or missing?
- What additional resources or training would have helped?
- What would you change to improve future annotation work?
- Any other thoughts or observations?

Assignment 3: Linguistic Annotation

- Submit your annotation files (.txt format) for Task 1 and Task 2.
- Each file is worth 1 points (2 points total).
- Make sure that you include your outputs from the POS/dep parsing.
- Due this Friday!

Midway presentation/report

Midway presentation and report

- More updates on Lamakū and course webpage.
- `https://hksung.github.io/Fall26_SLS775/assignments/3_midway`
- `https://hksung.github.io/Fall26_SLS775/assignments/4_midway2`