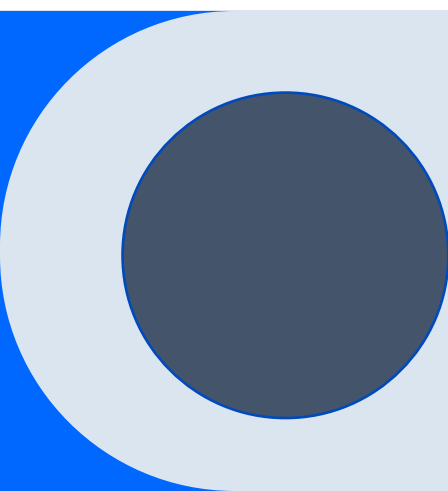


Annotation schema and template for spoken and written corpus

Athanasios Karasimos
(Aristotle University of Thessaloniki)



Workshop on Corpus-linguistic Applications
Foreign Language Teaching Laboratory (ENL AUTH)

Friday, 10 February 2023



Outline

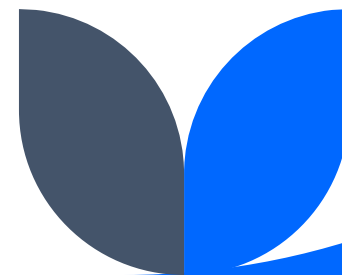
Introduction

Schema vs. template

Why Annotation?

Spoken vs. written corpus annotations: best practices and guidelines

Creating an annotation template (Praat, ELAN, CatMa)



Introduction

Processing Text and speech: why?

Data and Metadata

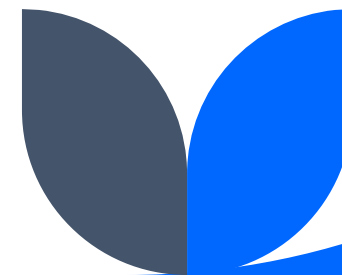
- 80% of information and texts are unstructured
- Full-information chains in organization texts
- Products and e-markets: presentations, advs, promotions
- Material from users: blogs, forums/ fora, wikis
- Customer opinions: social media, personal analysis

Enormous amount of data

- 161,000,000 GBs in digital content in 2006
- ~ 1000 EBs in digital content in 2010. In 2022???
- Sound and image need abstracts and labels
- Numerous text bodies without annotation and metadata

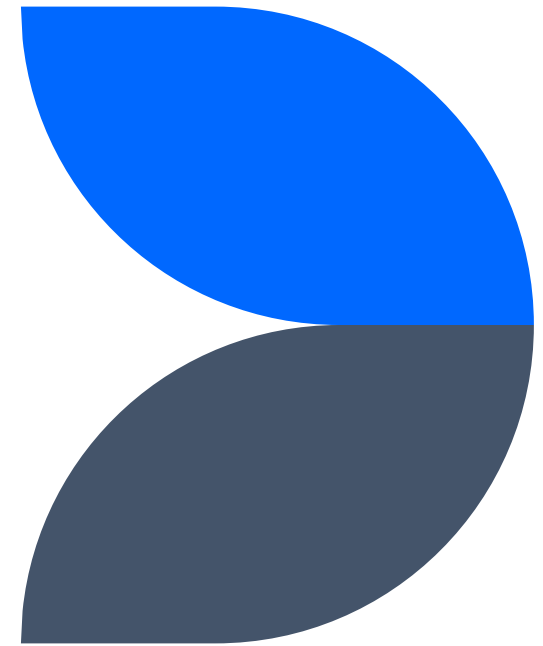
Ambiguity

A perhaps surprising fact about these categories of linguistic knowledge is that most tasks in speech and language processing can be viewed as resolving **ambiguity** at one of these levels.



Scheme vs. template

From data and metadata to
annotation



What is a scheme?

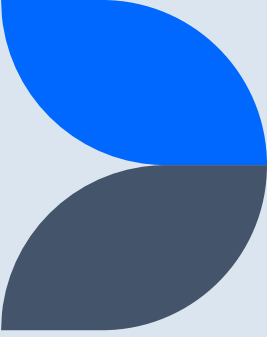
An annotation scheme is a set of rules used to determine how a document or other digital object is annotated. It includes guidelines for how the annotations should be formatted and includes the tags and attributes that should be used. It is used by annotation programs or human annotators to ensure consistency in the annotations across multiple documents.

What is a template?

An annotation template is a document used to create an annotation for a specific item, such as a book, article, website, or other analogical or digital resource. It includes fields for summarizing the content and evaluating its relevance, accuracy, and quality. It usually includes a section to add any personal comments or insights.

The template comes before the scheme.

Corpus annotation

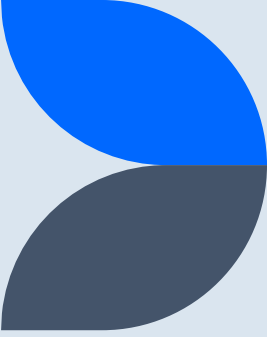


Corpus annotation is the practice of adding interpretative linguistic information to a corpus.

For example, one common type of annotation is the addition of tags, or labels, indicating the word class to which words in a text belong.

This is so-called part-of-speech tagging (or POS tagging), and can be useful, for example, in distinguishing words which have the same spelling, but different meanings or pronunciation. If a word in a text is spelt present, it may be a noun (= 'gift'), a verb (= 'give someone a present') or an adjective (= 'not absent').

Corpus annotation



The meanings of these same-looking words are very different, and also there is a difference of pronunciation.

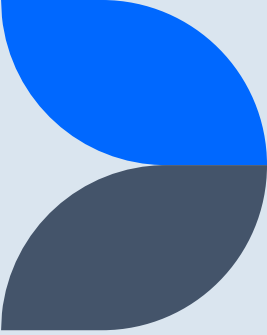
Using one simple method of representing the POS tags, these three words may be annotated as follows:

present_NN1 (singular common noun)

present_VVB (base form of a lexical verb)

present_JJ (general adjective)

Corpus annotation



Some people prefer not to engage in corpus annotation: **the unannotated corpus is the 'pure' corpus they want to investigate** — the corpus without adulteration with information which is suspect, possibly reflecting the predilections, or even the errors, of the annotator.

For others, annotation is a means to make a corpus much more useful — **an enrichment of the original raw corpus**. From this perspective, probably a majority view, adding annotation to a corpus is giving 'added value', which can be used for research by the individual or team that carried out the annotation, but which can also be passed on to others who may find it useful for their own purposes.

For example, POS-tagged versions of major English language corpora such as the Brown Corpus, the LOB Corpus and the British National Corpus have been distributed widely throughout the world.

“

What different kinds of
annotation are there?

Geoffrey Leech

”

Phonetic and syntactic annotation

Phonetic annotation - adding information about how a word in a spoken corpus was pronounced.

prosodic annotation — again in a spoken corpus — adding information about prosodic features such as stress, intonation and pauses.

syntactic annotation — e.g. adding information about how a given sentence is parsed, in terms of syntactic analysis into such units such phrases and clauses

Semantic annotation

adding information about the semantic category of words — the noun cricket as a term for a sport and as a term for an insect belong to different semantic categories, although there is no difference in spelling or pronunciation.

The ontology aspect

The thematic roles annotation (FRAMEnet)

Activity_prepare

Definition:

An **Agent** prepares for an **Activity**.

The troops were **PREPARING** **themselves** for the mission.

The boats are **GETTING READY** to leave the pier.

FEs:

Core:

Activity [Act]

This FE identifies the **Activity** for which an **Agent** is preparing.

Agent [Agent]

Semantic Type: Sentient
Non-Core:

An **Agent** prepares for an **Activity**.

Beneficiary []

This extrathematic FE applies to participants that derive a benefit from the occurrence of

Degree []

Semantic Type: Degree
Depictive [Depict]

This FE describes the amount of preparation the Agent puts into an activity.

This FE is used for the **Depictive** phrase describing the actor or undergoer of an action.

Duration [Dur]

Semantic Type: Duration
Event_description []

This FE identifies the length of **Time** during which an **Agent** prepares for an **Activity**.

This FE describes the state of affairs denoted by the target as role fillers in other frames.

Iterations [Itc]

The frame element **Iterations** is used for expressions that indicate the number of times an

Manner [Mnr]

Semantic Type: Manner
Means [Means]

Semantic Type: State_of_affairs
Place [Place]

Semantic Type: Locative_relation
Purpose [Purp]

Semantic Type: State_of_affairs
Time [Time]

Semantic Type: Time

This FE identifies the **Manner** in which an **Agent** prepares for an **Activity**.

This FE identifies the **Means** with which an **Agent** prepares for an **Activity**.

This FE identifies the **Place** where an **Agent** prepares for an **Activity**.

This FE identifies the **Purpose** for which an **Agent** prepares for an **Activity**.

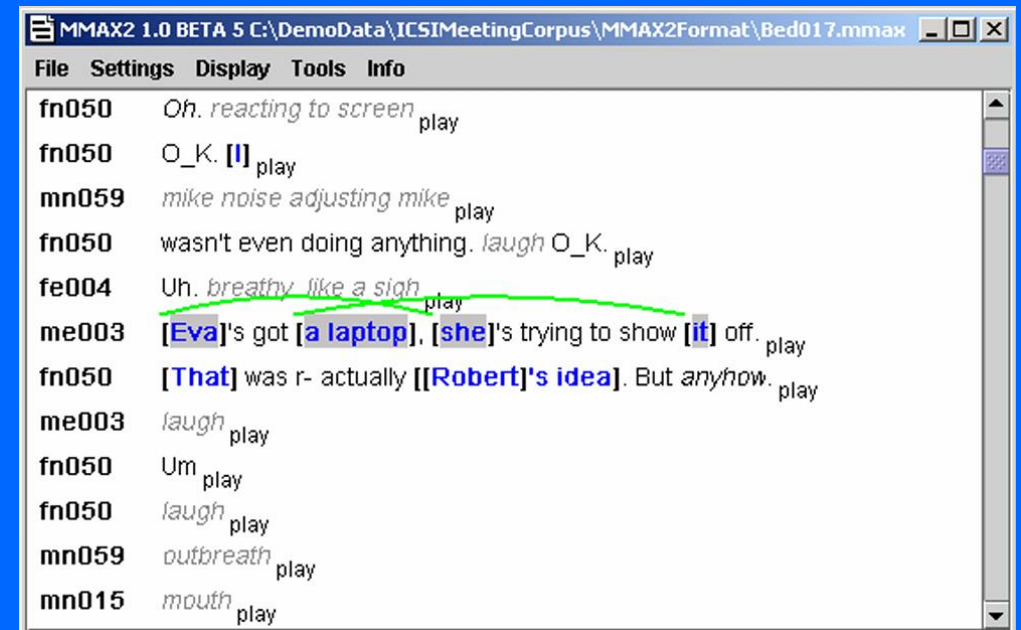
This FE identifies the **Time** at which an **Agent** prepares for an **Activity**.

Pragmatic annotation

adding information about the kinds of speech act (or dialogue act) that occur in a spoken dialogue — thus the utterance *okay* on different occasions may be an acknowledgement, a request for feedback, an acceptance, or a pragmatic marker initiating a new phase of discussion.

Discourse annotation

adding information about anaphoric links in a text, for example connecting the pronoun them and its antecedent the horses in: I'll saddle the horses and bring them round. [an example from the Brown corpus]

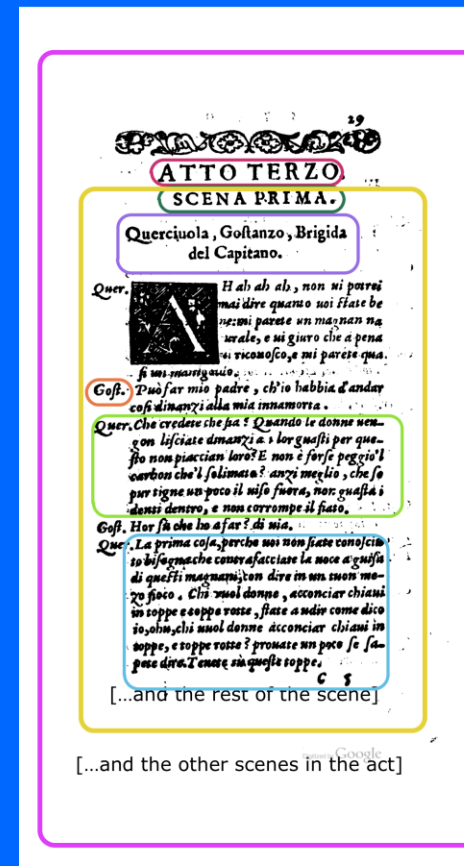


Stylistic annotation

adding information about speech and thought presentation (direct speech, indirect speech, free indirect thought, etc.)

Authorship tracking

Drafting and “track changes”



```
<div type="act">
<head>Atto Terzo</head>
<div type="scene">
<head>Scena Prima</head>
<stage>Querciuiola, Costanzo, Brigida
<lb/>del Capitano</stage>
<sp><speaker>Quer.</speaker>
<p>Ah ah ah ah, non ui potrei mai dire quanto uoi state bene: me parete un magnan naturale, e ui giuro che a pena riconosco, e mi parete quasi un <gap/>.</p>
</sp>
<sp><speaker>Gost.</speaker>
<p>Può far mio padre, ch'io habbia d'andar cosi dinanzi alla mia innamorata.</p>
</sp>
<sp><speaker>Quer.</speaker>
<p>Che credete che sia? Quando le donne uengon lasciate dinanzia lorquasti per questo non piaccian loro? E non è forse peggio'l carbon che'l solimata? anzi meglio, che se pur tigne un poco il uiso fuera, non guasta i denti dentro, e non corrompe il fiato.</p>
</sp>
<sp><speaker>Gost.</speaker>
<p>Hor su che ho afar? di uia.</p>
</sp>
<sp><speaker>Quer.</speaker>
<p>La prima cosa, perche uoi non siate conosciuto bisognate contrasacciate la uoce a guisa di questi magnani, con dire in un tuon mezzo fioco. Chi vuol donne, acconciar chiaui in toppe, e toppe rotte, state a udire come dico io, oh, chi vuol donne acconciar chiaui in toppe, e toppe rotte? prouate un poco se sapete dire. Tenete su queste toppe.</p>
</sp>
[...and the rest of the scene]
</div>
[...and the other scenes in the act]
</div>
```

Lexical annotation

adding the identity of the lemma of each word form in a text — i.e. the base form of the word, such as would occur as its headword in a dictionary (e.g. *lying* has the lemma LIE).

The wordlists and the lemma-lists

Why annotate?

1

Manual examination of a corpus

What has been built into the corpus in the form of annotations can also be extracted from the corpus again, and used in various ways

2

Automatic analysis of a corpus

i.e. corpora which have been POS-tagged can automatically yield frequency lists or frequency dictionaries with grammatical classification.

3

Re-usability of annotations

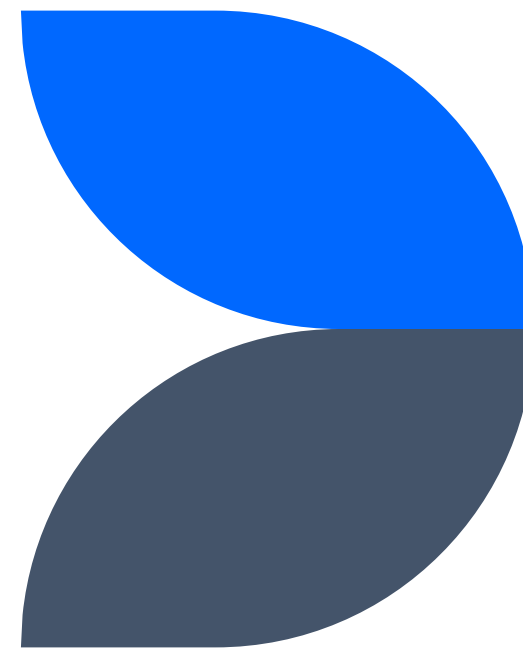
This argument may work for some cases, but generally the annotation is far more useful if it is preserved for future use.

4

Multi-functionality

If we take the re-usability argument one step further, we note that annotation often has many different purposes or applications: it is **multi-functional**.

Useful standards for corpus annotation



Annotations should be separable

The annotations are added as an 'optional extra' to the corpus. It should always be easy to separate the annotations from the raw corpus, so that the raw corpus can be retrieved exactly in the form it had before the annotations were added. This is common sense: not all users will find the annotations useful, and annotation should never result in any loss of information about the original corpus data.



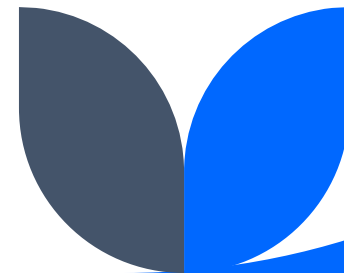
Detailed & explicit documentation

How, where, when and by whom were the annotations applied?

Mention any computer tools used, and any phases of revision resulting in new releases, etc.

What annotation scheme was applied?

An annotation scheme is an explanatory system supplying information about the annotation practices followed, and the explicit interpretation, in terms of linguistic terminology and analysis, for the annotation.



Behind-the-scenes: typology

Any type of annotation presupposes a typology — a system of classification — for the phenomena being represented.

But linguistics, like most academic disciplines, is sadly lacking in agreement about the categories to be used in such description.

Different terminologies abound, and even the use of a single term, such as verb phrase, is notoriously a prey to competing theories. Even an apparently simple matter, such as defining word classes (POS), is open to considerable disagreement.

There is no absolute 'God's truth' view of language or 'gold standard' annotation against which the decision to call word *x* as noun and word *y* a verb can be measured.

FRIDA typology (Granger, 2003)

COPLE2 corpus

Token value (w-175): novidades

XML	Raw XML value	<code>nov<del hand="corrector">e<add hand="corrector">i</add</code>
form	Student form	novidades
fform	Corrected form	novidades
nform	Normalized form	novidades
<hr/>		
pos	POS tag	[select] ▼
mfs	Morphological features	fp
lemma	Lemma	novidade
error	Error code(s)	[select] ▼

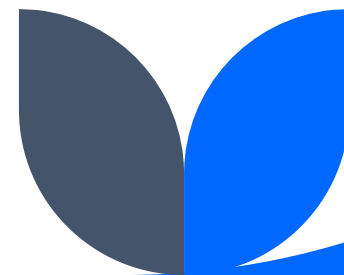
Error domain		Error categories	
<F>	Form	<AGL> <MAJ> <DIA> <HOM> <GRA>	Agglutination Upper/lower case Diacritics Homonymy Other spelling errors
<M>	Morphology	<MDP> <MDS> <MFL> <MFC> <MCO>	Derivation-prefixation Derivation-suffixation Inflection Inflection confusion Compounding
<G>	Grammar	<CLA> <AUX> <GEN> <MOD> <NBR> <PER> <TPS> <VOI> <EUF>	Class Auxiliary Gender Mode Number Person Tense Voice Euphony
<L>	Lexis	<SIG> <CPA> <CPD> <CPV> <CPN> <FIG>	Meaning Adjective complementation Adverb complementation Verb complementation Noun complementation Prefab
<X>	Syntax	<ORD> <MAN> <RED> <COH>	Word order Word missing Word redundant Cohesion
<R>	Register	<RLE> <RSY>	Lexis Syntax
<Y>	Style	<CLR> <LOU>	Unclear Heavy
<Q>	Punctuation	<CON> <TRO> <OUB>	Punctuation confusion Punctuation redundant Punctuation missing
<Z>	Typo		

Annotation practices and de facto standards I

By de facto standards, I mean some kind of standardisation that has already begun to take place, due to influential precedents or practical initiatives in the research community.

These contrast with de iure or 'God's truth' standards do not exist.

'God's truth' standards, if they existed, would be imposed from on high. De facto standards, on the other hand, emerge (often gradually) from the research community in a bottom-up manner.

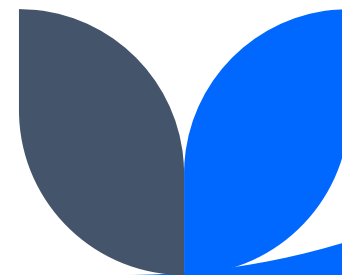


Annotation practices and de facto standards II

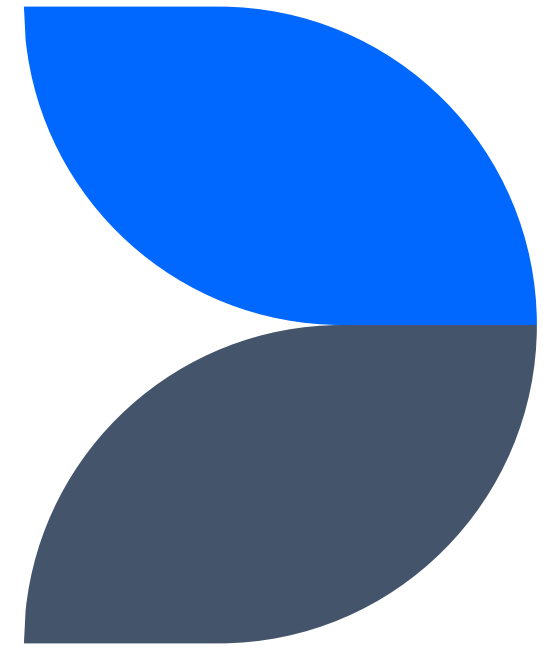
De facto standards encapsulate what people have found to work in the past, which argues that they should be adopted by people undertaking a new research project, to support a growing consensus in the community.

A new project breaks new ground, for example with a different kind of data, a different language, a different purpose those of previous projects.

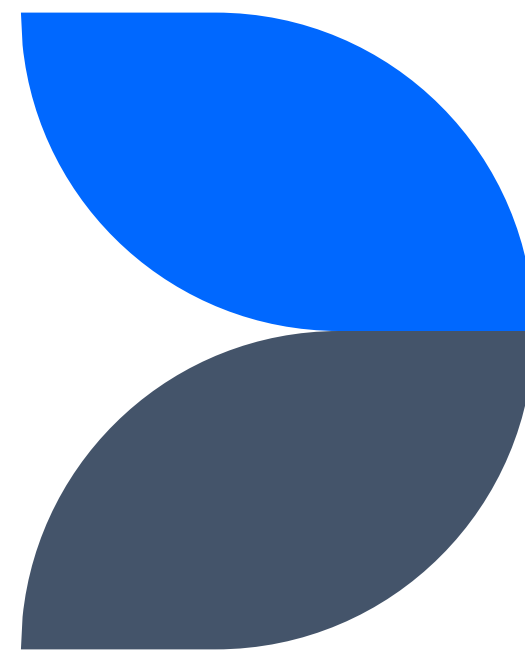
Nevertheless, it makes sense for new projects to respect the outcomes of earlier projects, and only to depart from their practices where this can be justified.



Encoding of Annotations



Manual annotation



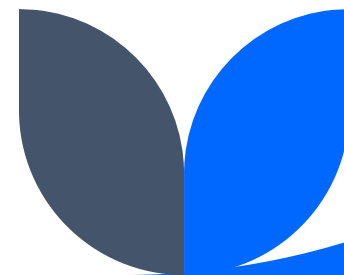
Annotations tags and explanation

This list acts as a glossary — a convenient first port of call for people trying to make sense of the annotations. For POS tagging, the first thing to list is the tagset — i.e., the list of symbols used for representing different POS categories. Such tagsets vary in size, from about 30 tags to about 270 tags. The tagset can be listed together with a simple definition and exemplification of what the tag means:

NN1 singular common noun (e.g. book, girl)

NN2 plural common noun (e.g. books, girls)

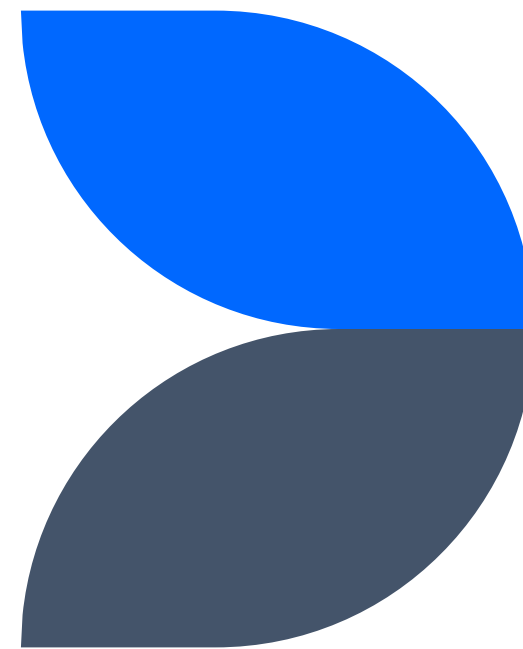
NP1 singular proper noun (e.g. Susan, Cairo)



A specification of annotation practices

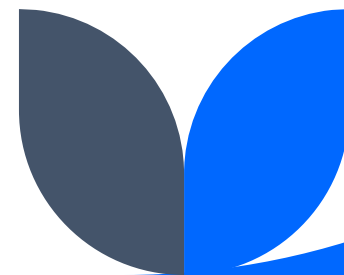
1. *segmentation*: e.g. assignment of POS tags assumes a prior segmentation of the corpus into words. This may involve 'grey areas' such as how to deal with hyphenated words, acronyms, enclitic forms such as the n't of don't.
2. *embedding*: e.g. in parsing, some units, such as words and phrases, may be included in other units, such as clauses and sentences; certain embeddings, however, may be disallowed. In effect, a grammar of the parsing scheme has to be supplied. Even POS tagging has to involve some embedding when we come to segment examples such as the New York-Los Angeles flight.
3. *the rules or guidelines for assigning particular annotation tags to text.*

Best practice for different linguistics levels



Part-of-speech (POS) tagging

- The 'Brown Family' of corpora (consisting of the Brown Corpus, the LOB Corpus, the Frown Corpus and the FLOB Corpus) makes use of a family of similar tagging practices, originated at Brown University and further developed at Lancaster. The two tagsets (C5 and C7) used for the tagging of the British National Corpus are well known (see Garside et al. 1997).
- An EAGLES document which recommends flexible 'standard' guidelines for EU languages is to be found in Leech and Wilson (1994), revised and abbreviated in Leech and Wilson (1999).
- Note that POS tagging schemes are often part of parsing schemes, to be considered under the next heading.



Syntactic annotation

- A well-developed parsing scheme already mentioned is that of the SUSANNE Corpus, Sampson (1995).
- The Penn Treebank and its accompanying parsing scheme has been the most influential of constituent structure schemes for syntax. (see Marcus et al 1993)
- Other schemes have adopted a dependency model rather than a constituent structure model — particularly the Constraint Grammar model of Karlsson et al. (1995).
- Leech, Barnett and Kahrel (1995) is another EAGLES 'standards-setting' document, this time focusing on guidelines for syntactic annotation. Because there can be fundamentally different models of syntactic analysis, this document is more tentative (even) than the Leech and Wilson one for POS tagging.



Prosodic annotation

- The standard system for annotating prosody (stress, intonation, etc.) is ToBI (= Tones and Break Indices), which comes with its own speech-processing platform. Its phonological model originated with Pierrehumbert (1980). The system is partially automated, but needs to be substantially adapted for fresh languages and dialects.
- ToBI is well supported by dedicated software and a committed research community. On the other hand, it has met with criticism, and two alternative annotation systems worth examining are INTSINT (see Hirst 1991) and TSM — tonetic stress marks (see Knowles et al. 1996).
- For a survey of prosodic annotation of dialogue, see Grice et al. (2000: 39-54).



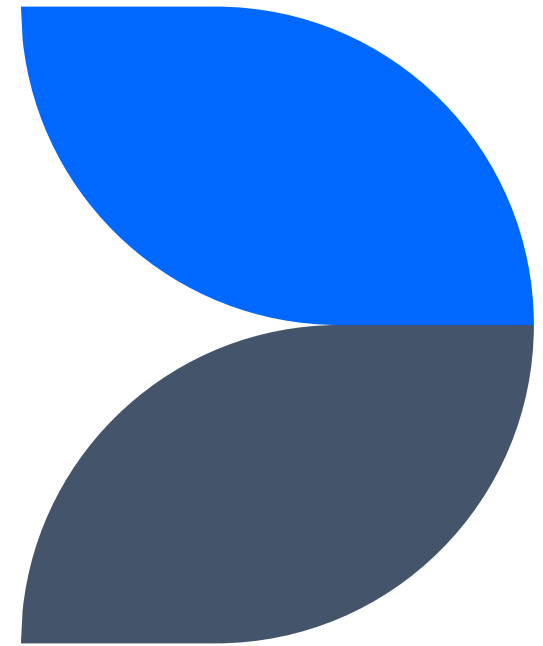
Pragmatic/Discourse annotation

- An international Discourse Resource Initiative (DRI) came up with some recommendations for the analysis of spoken discourse at the level of dialogue acts (= speech acts) and at higher levels such as dialogue transactions, constituting a kind of 'grammar' of discourse. These were set out in the DAMSL manual (= Dialog Act Markup in Several Layers) (Allen and Core 1997).
- Other influential schemes are those of TRAINS, VERBMOBIL, the Edinburgh Map Task Corpus, SPAAC (Leech and Weisser 2003). These all focus on practical task-oriented dialogue. One exceptional case is the Switchboard DAMSL annotation project (Stolcke et al. 2000), applied to telephone conversational data.
- Discourse can also be analysed at the level of anaphoric relations (e.g. pronouns and their antecedents — see Garside et al 1997:66-84).
- A survey of pragmatic annotation is provided in Grice et al. (2000: 54-67).
- A European project MATE (= Multi-level annotation, tools engineering) has tackled the issue of standardization in developing tools for corpus annotation, and more specifically for dialogue annotation, developing a workbench and an evaluation of various schemes, investigating their applicability across languages (<http://mate.nis.sdu.dk/>).



Evaluation of annotation

realism, accuracy and consistency



Realism, accuracy and consistency

The quality or 'goodness' of annotation was one important — though rather unclear — criterion to be sought for in annotation. Reverting to the POS-tagging example once again, we may distinguish two quite different ideas of quality. The first refers to the linguistic realism of the categories. A notion of quality refers not to the tagset, but to the accuracy and consistency with which it is applied.



Questions about evaluation I

What is meant by 'correct'?

The answer is: 'correctness' is defined by what the annotation scheme allows or disallows — and this is an added reason why the annotation scheme has to be specific in detail, and has to correspond as closely as possible with linguistic realities recognized as such.



Questions about evaluation II

Is it possible for hand-editors to achieve 100% accuracy?

Most people will find this unlikely, because of the unpredictable peculiarities of language that crop up in a corpus, and because of the failure of even the most detailed annotation schemes to deal with all eventualities. Perhaps between 99% and 99.5% accuracy might be the best that can be achieved, given that unclear and unprecedented cases are bound to arise.



Questions about evaluation III

How consistently has the annotation task been performed?

One way to test this in POS tagging is to have two human annotators post-edit the same piece of automatically-tagged text, and to determine in what percentage of cases they agree with one another. The more this consistency measure (called inter-rater agreement) approaches 100%, the higher the quality of the annotation. (Accuracy and consistency are obviously related: if both raters achieve 100% accuracy, it is inevitable that they achieve 100% consistency.)



The practical task of annotation

It is useful to say something about the practicalities of corpus annotation. Assume, say, that you have a text or a corpus you want to work on, and want to 'get the tags into the text'.

The practical task of annotation

It is not necessary to have special software. You can annotate the text using a general-purpose text editor or word processor. But this means the job has to be done by hand, which risks being slow and prone to error.

For some purposes, particularly if the corpus is large and is to be made available for general use, it is important to have the annotation validated. That is, the vocabulary of annotation is controlled and is allowed to occur only in syntactically valid ways. A validating tool can be written from scratch, or can use macros for word processors or editors.

If you decide to use XML-compliant annotation, this means that you have the option to make use of the increasingly available XML editors. An XML editor, in conjunction with a DTD or schema, can do the job of enforcing well-formedness or validity without any programming of the software, although a high degree of expertise with XML will come in useful.

The practical task of annotation

Special tagging software has been developed for large projects — for example the CLAWS tagger and Template Tagger used for the Brown Family or corpora and the BNC. Such programs or packages can be licensed for your own annotation work. (For CLAWS, see the UCREL website <http://www.comp.lancs.ac.uk/ucrel/>.)

There are tagsets which come with specific software — e.g. the C5, C7 and C8 tagsets for CLAWS, and CHAT for the CHILDES system, which is the de facto standard for language acquisition data. There are several natural language processing tools for multi-linguistic annotations.

There are more general architectures for handling texts, language data and software systems for building and annotation corpora. For example, a prominent example of this is GATE ('general architecture for text engineering' <http://gate.ac.uk>) developed at the University of Sheffield.

The template samples

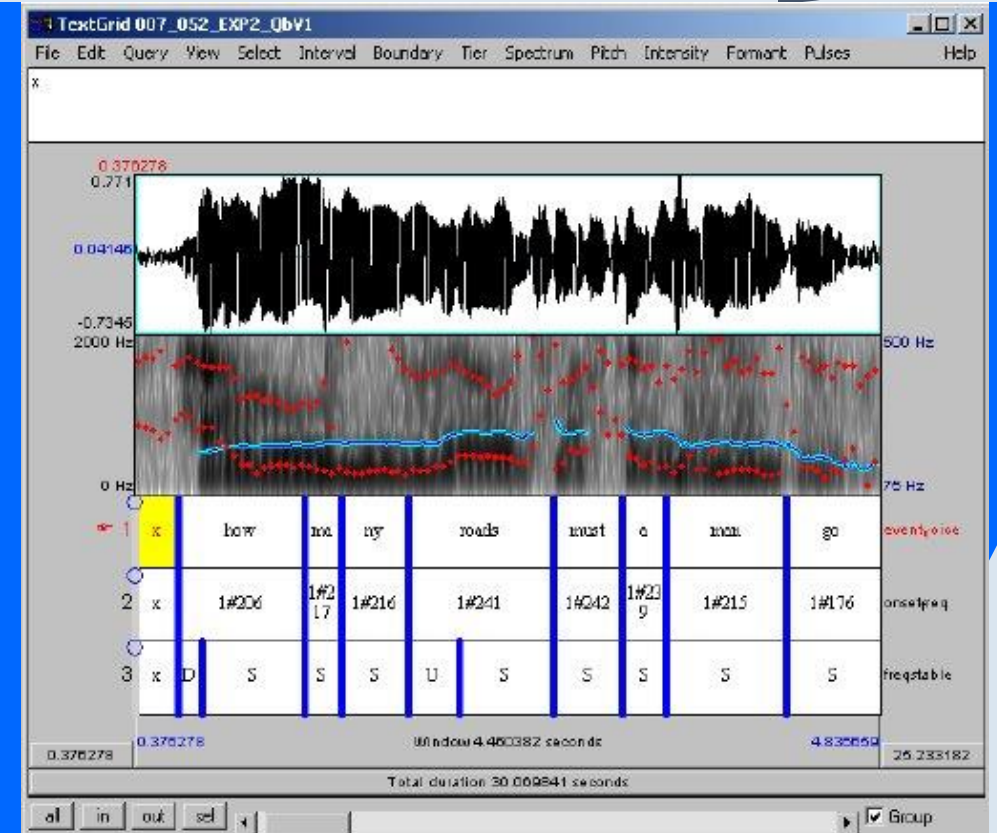
Praat (spoken data)

ELAN (spoken data [audio and video])

CatMa (written data)

Praat annotations

- Independent tiers (clones tiers, no-multi info tiers, no controlled vocs)
- Orthographic transcription vs. phonetic (IPA) transcription
- Third-party tools (Transcriptor, S2T converter)
- Praat scripts



CatMa Annotations

- Tagsets, tags and subtags
- Tag properties
- Annotation template and user history
- Annotations extractions, stats, analysis

The screenshot displays the CATMA 6.5.4 interface. The top bar shows 'CATMA 6.5.4' and 'DELEAR Corpus'. The main window is titled '2020-2021 102 FINAL 146'. The left sidebar has buttons for 'Project', 'Tags', 'Annotate', and 'Analyze'. The central pane shows a text document with various colored annotations (red, green, blue, purple) and a vertical toolbar on the left. The right pane shows 'Collection currently being edited' and '2020-2021 102 Final 146 Default Annotations'. Below this is a 'Tagsets' table and a 'Selected Annotations' table.

Tagsets	Tags	Properties	Values
Punctuation	▶ CON,PMI		👁
Form	▶ CAP,HOM,SPE...		👁
Grammar			👁
	ADV		👁
	AUX		👁
	CAS		👁
	CONJ		👁
	DET	▶ DET_OVU,D	👁
	DET_MIS		👁
	DET_OMI		👁
	DET_OVU		👁
	GEN		👁
	MOV		👁

Selected Annotations				
Annotation	Tag	Author	Collection	Tagset
at	PRE	Federica_1997	2020-202...	Grammar

Thank you

Athanasios Karasimos

akarasimos@enl.auth.gr |

akarasimos@gmail.com

Aristotle University of Thessaloniki