Text Models

Raf Alvarado
UVA DS 5001

Covers hermeneutics, the OHCO model of text, and mapping these models to dataframes

Business

Homework 1

- Two videos added
- Extension to Wednesday (midnight)
- 4 questions (removed the first)
- Submit PDFs

Rivanna

- Allocation: `msds_ds5001`

Office hours

- Fridays at 2:00 PM on Zoom

Review

Crane's Challenge

"What do you do with a million books?"

Describes changes to reading that come with **digitization at scale**

We go from **reading texts** to **mining culture** from corpora

Examples

Culturomics — n-grams

Cultural Analytics — visualizing text at scale

Distant Reading — interpretation of corpus-level patterns

Review

ETA studies “**text as text**”

- Requires **domain knowledge** of what a text is

Text has properties **distinct from language** as grammar

- *langue* vs *parole* / grammar vs discourse

Some **properties** of text as text:

- Language **produces** texts (language is latent)
- Text provide the **data** by which we study language
- Texts “contain” patterns of **culture**, symbolic structures
- Texts can reveal **social** patterns (as social sensors)
- → Texts have **structures and functions** that may be modeled

Course in General Linguistics

Ferdinand de Saussure

Translated by Wade Baskin

Edited by Perry Meisel and Haun Saussy

1916

ETA Overview

ex: just looking at the tokens, we can see the difference between words that do grammatical work and words that do other things

Order of knowledge

Text → Language → Culture → History

Text models

Text are the data on which everything else is built

From trees to tables

Language models

Derived from textual corpora

Texts must be modeled first (often overlooked)

Culture and History models

Also derived from corpora + language models

Today's readings focus on
properties of text from domain experts

Hermeneutics and **text encoding**
both provide **models of text**

Text models usually **missing**
from discussions of text mining, NLP, etc.

Ricoeur

The Ricoeur reading for this week stands for the field of **hermeneutics**

> The reading provides a general theory of the **structure** and **function** of texts

Structure

> Texts are shaped by their being a kind of **discourse**, a form of locution that is **fixed** and **distanciated** by writing

Function

> Texts make possible both the **existence** of society and its **study** (in text-based societies)

**exception is the Inca society

The Model of the Text: Meaningful Action Considered as a Text

see: the unfamiliar words

Paul Ricœur

MY AIM IN THIS PAPER will be to test an hypothesis. I assume that the primary sense of the word "hermeneutics" concerns the rules required for the interpretation of the written documents of our culture. In assuming this starting point I am remaining faithful to the concept of *Auslegung* as it was stated by Wilhelm Dilthey; whereas *Verstehen* (understanding, comprehension) relies on the recognition of what a foreign subject means or intends on the basis of all kinds of signs in which psychic life expresses itself *(Lebensäusserungen)*, *Auslegung* (interpretation, exegesis) implies something more specific: it covers only a limited category of signs, those which are fixed by writing, including all the sorts of documents and monuments which entail a fixation similar to writing.

Guide for the Perplexed

Ricoeur's writing is a bit hard to read if you are not familiar with **philosophical writing**

Uses lots of keywords, many in **German**
Cites **thinkers** and **ideas** you may never have heard of or read

This is because **Western philosophy** intentionally builds on previous ideas

It is **a long dialog** that goes back to **Plato**

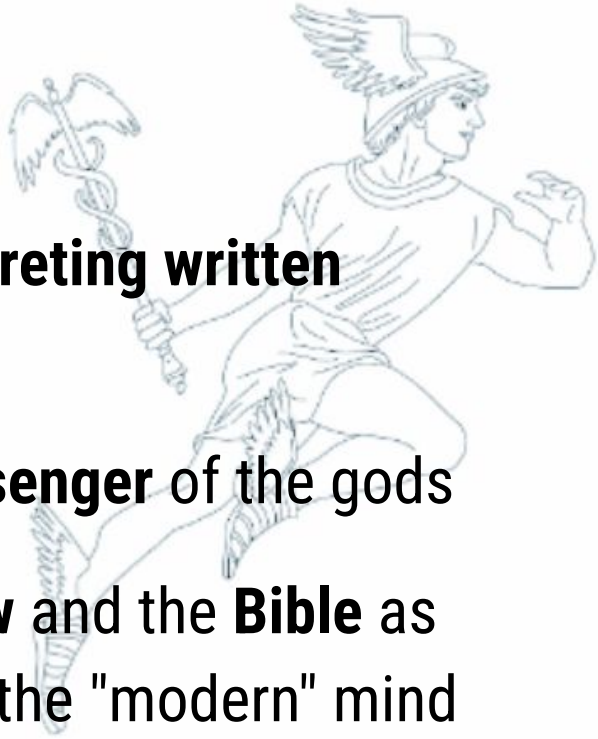
(Made possible by texts and **textual culture)**

This lesson provides some **background information** to help you understand the text

Coincidentally, a quintessential hermeneutic task : -)

Hermeneutics

interpretation theory***



Hermeneutics is the theory and practice of **interpreting written documents** in our culture

- Its lexical root is the Greek god *Hermes*, **messenger** of the gods
- It began with the need to interpret **Roman Law** and the **Bible** as these documents become more **distant** from the "modern" mind

Because of the fundamental role of **language** in society and culture, the idea of interpretation has been used to distinguish the natural from the human sciences

- Natural science → **explanation**
- Human science → **understanding** → related to **interpretation**

can't study human beings without language

langage:human science::math:natural science

***motives, description of what happened and draw parallels

Hermeneutics is German

Hermeneutics began in Germany and was developed by German philosophy. Here are some **keywords**:

Wissenschaft	Knowledge, science
Geist	Spirit, mind, intellect (literally “ghost”)
Geisteswissenschaft	The human sciences
Naturwissenschaft	The natural sciences
Verstehen	Understanding, comprehension
Auslegung	Interpretation, exegesis (*ulema*)
Erklären	Explanation, use of logic
Welt	World
Umwelt	Environment, surroundings, situation

the situation in which you speak

Interpretive Social Science

Wilhelm **Dilthey**, an influential 19th-century German philosopher, expanded hermeneutics to be a general method to study humans and society

> In contrast to positivist approaches, the goal of social science is to **interpret human behavior**, to **understand** it, not to reduce it to causal or structural **explanation**
>
> This is in contrast to **positivism**, which is the idea that only measurable behavior matters -- i.e. **statistics**

It is easy to dismiss this view, but **human behavior remains difficult to explain** in positivist terms

Verstehen

Central to this idea is the concept of **verstehen**, *understanding*

- Human **behavior** is understood to be **mediated** by verstehen — i.e. ethos (values) and worldview

The debate continues to this day regarding **the role of culture** in human behavior material vs. culture, what influences our beliefs

- Recently, **behavioral economics** has come to understand so-called “non-rational behavior”
- But culture is often reduced to forms of **bias** . . .
- Culture is more complicated and interesting than this!

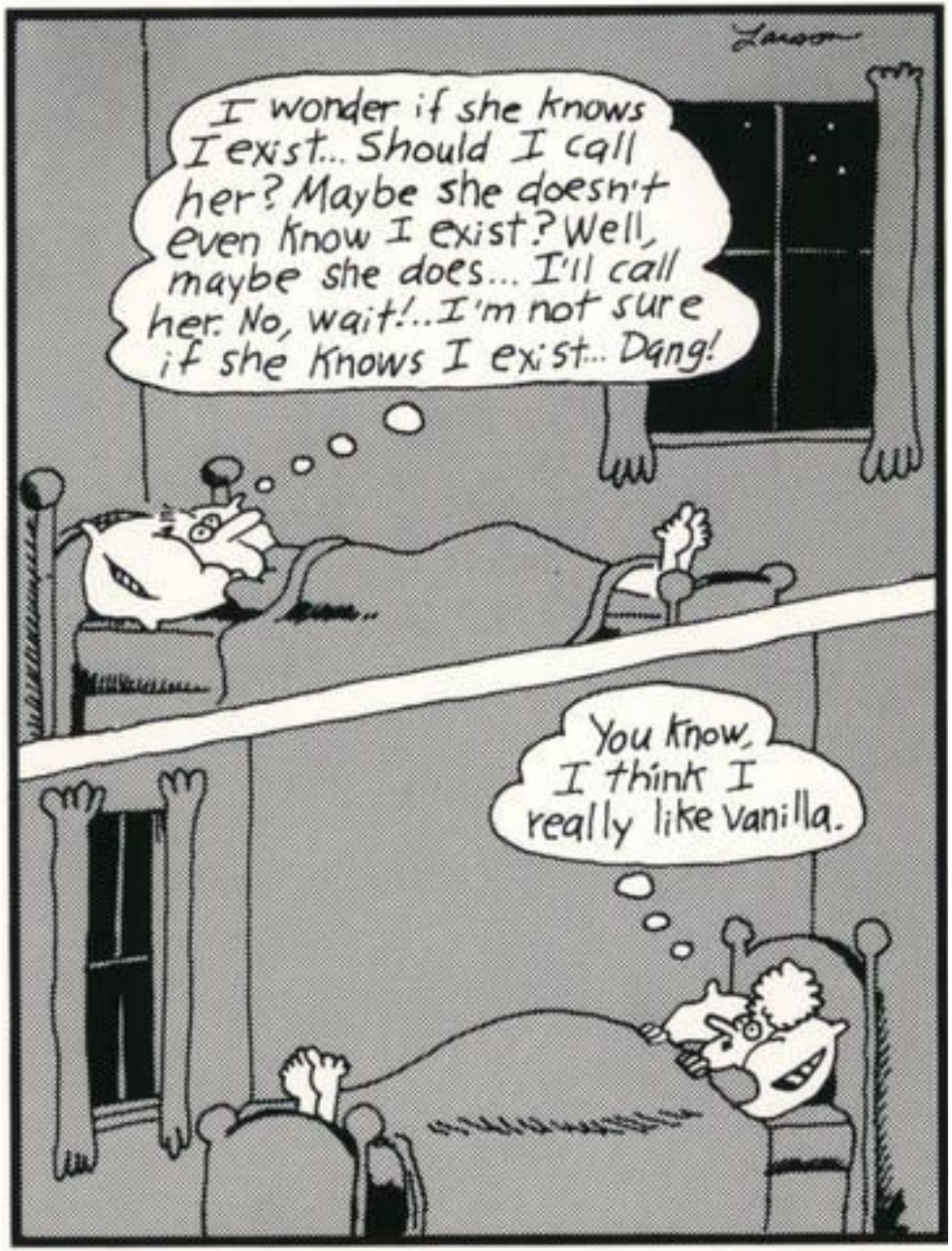
To sum it up:

Humans inhabit **worlds**, not just **situations**

Humans ***act***, not merely ***react*** (behave)

*** anti behaviorist views

Humans engage in **"meaningful action"**



Same planet, different worlds

So, what does this have to do with text?

Texts have specific **properties** with a special **relationship to human action**

The methods used to **interpret texts** allow us to study **society and culture**

This is Ricoeur's **hypothesis**

Unpacking Ricoeur

So, let's **unpack** some of the ideas presented in Ricoeur's essay, "The Model of the Text"

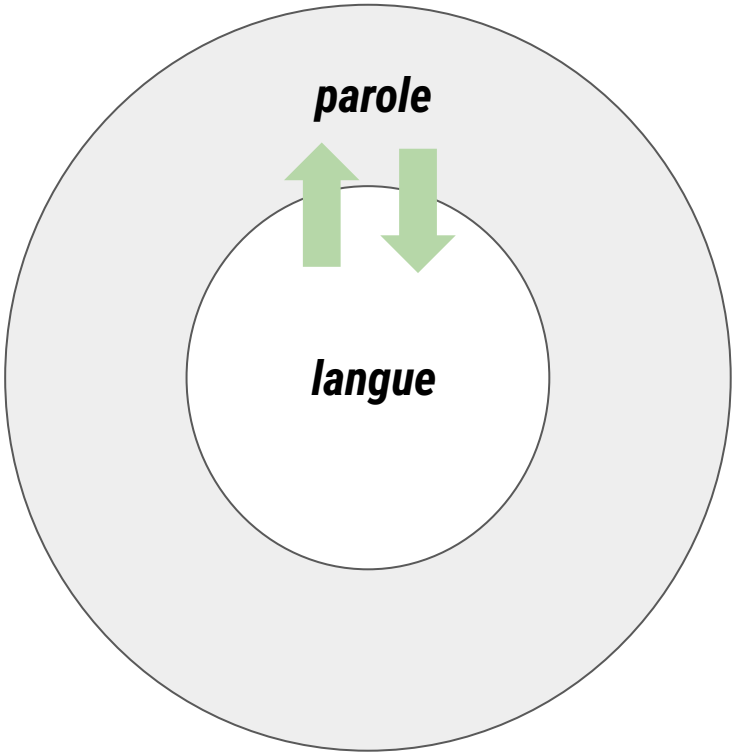
> In this essay, Ricoeur presents **a model of text** that is meant to provide the **foundation for the study of culture**
>
> This is because text for many societies is a major vehicle of **cultural memory**

Our concern is to summarize **the fundamental features of text** and to frame our analytical work

> We will present these as a series of **assertions** . . .

1. Text is not language

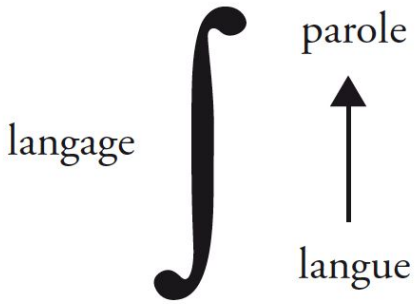
Saussure's Principle



The linguist Ferdinand Saussure distinguished between *langue* and *parole*, or **language** and **discourse**

Discourse (or speech) is what a person **actually says** — the words that are produced

Language (or grammar) refers to the system of rules that make discourse possible



Some properties of *Langue* and *Parole*

Language (*langue*)	Speech (*parole*)
Grammar	***Discourse***
Competence	Performance
Finite rules	Indefinite patterns
System	Usage
Collective	Individual
Unconscious	Conscious
“Structure”	**Event**
Latent	**Observed**

NLP

Text Mining

2. Discourse has its own linguistics

The Linguistics of Discourse

Discourse has ***structure*** and ***function***

Structurally, discourse consists of **sentences**, **paragraphs**, **dialogs**, **messages**, etc.

**not the word, the SENTENCE

The basic unit of discourse is the **sentence**

Functionally, discourse has **psychological** and **sociological motivations** and **effects**

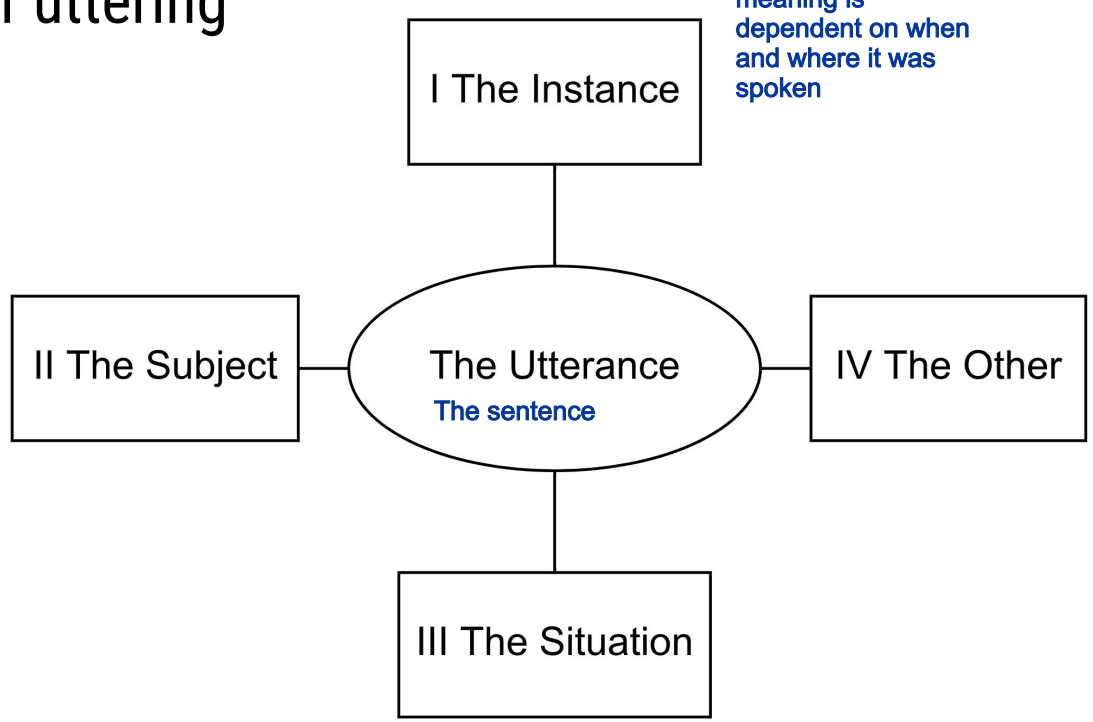
not just about neurons hahahha

The use of language (discourse) is bound up with human behavior and therefore cannot be studied apart from it

This is why the study of text is so important to the humanities and human sciences

And why Ricoeur and Dilthey found their work on it

Ricoeur identifies **four traits** associated with the **event** of uttering sentences



The place and time of the utterance

meaning is dependent on when and where it was spoken

The person making utterance; the intention; the sender

The person(s) to whom the utterance is addressed; the receiver

The external things shared by II and IV

in order to understand what's being said, you have to infer these things or they're given

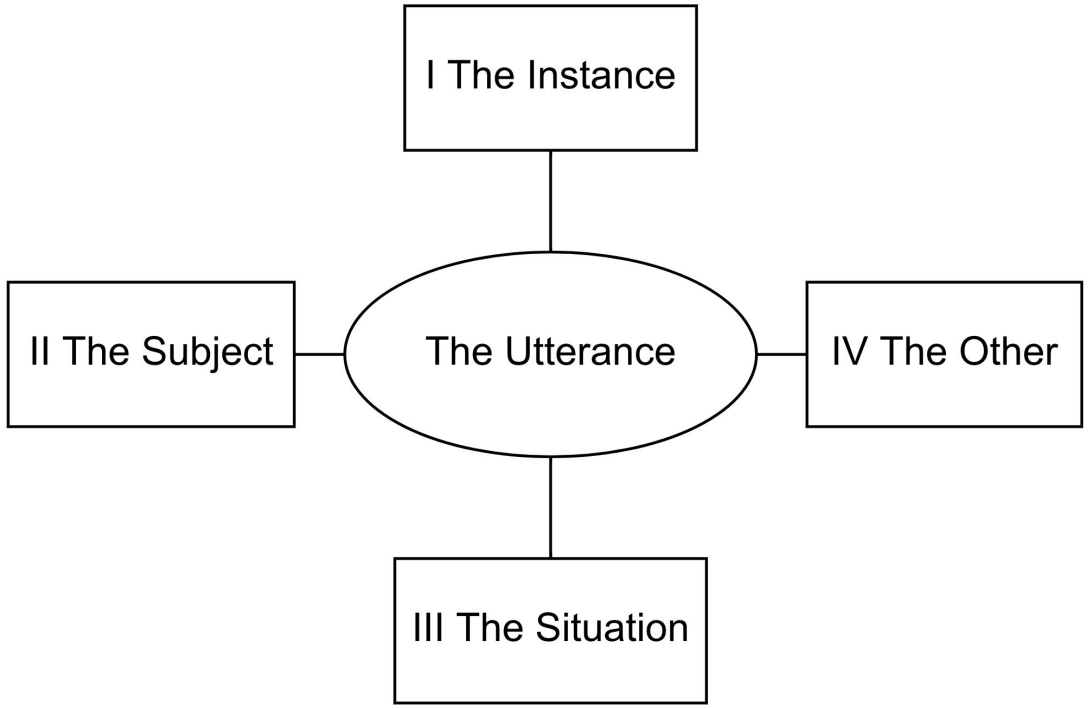
First trait: Discourse is always realized temporally and in a present, whereas the language system is virtual and outside of time. Emile Benvéniste calls this the "instance of discourse."

Second trait: Whereas language lacks a subject—in the sense that the question "Who is speaking?" does not apply at its level—discourse refers to its speaker by means of a complex set of indicators such as the personal pronouns. We will say that the "instance of discourse" is self-referential.

Third trait: Whereas the signs in language refer only to other signs within the same system, and whereas language therefore lacks a world just as it lacks temporality and subjectivity, discourse is always about something. It refers to a world which it claims to describe, to express, or to represent. It is in discourse that the symbolic function of language is actualized.

Fourth trait: Whereas language is only the condition for communication, for which it provides the codes, it is in discourse that all messages are exchanged. In this sense, discourse alone has not only a world, but an *other*—another person, an interlocutor to whom it is addressed.

Overlaying this structure are the various **aspects of meaning** as defined by **speech-act** theory (Austin, Searle, et al)



ILLOCUTION
The **force** or **sentiment** behind the utterance

thinking about motive and tone of how the sentence is said (emojis hehehhe, punctuation gives hints)

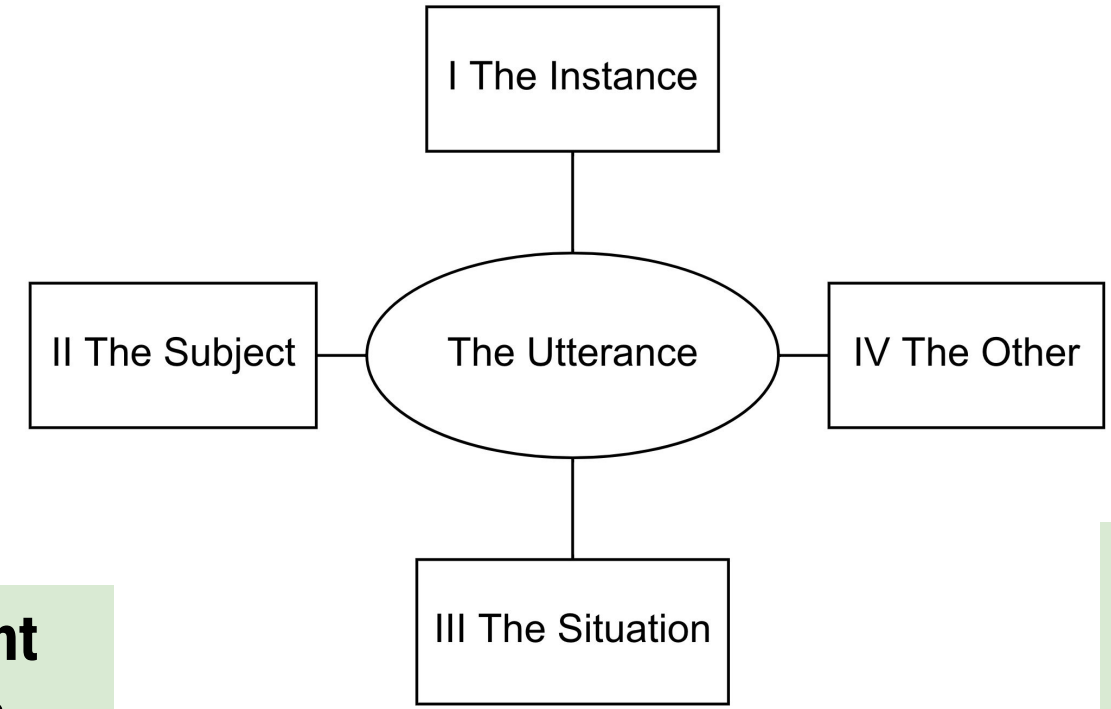
PERLOCUTION
The **effect** of the utterance on other people

the effect of language on ppl's behavior (politicians, marketers)

LOCUTION
The **content** of the utterance

what you're talking about

This structure also provides a map for much of **text analytics** . . .



usually the most important; not interested in what's being said, but HOW it's being said

***less defined domain

ILLOCUTION
The force or sentiment behind the utterance

PERLOCUTION
The effect of the utterance on other people

LOCUTION
The content of the utterance

Sentiment analysis, Stylistics

Event detection, A/B testing

Topic modeling

3. Only sentences have meaning

not words, example: can only extract the meaning of a words based on how it was used in a sentence (can't just go to a dictionary to get the meaning, because there are so many 'meanings' recorded

a sentence ASSIGNS a meaning to a word :)

Sentences and Meaning

Sentences have meaning in that they **assert something about something**

- **Sentences** implicitly have the form **A is B**
  - "Is" = the copula = an **intentional connecting** of ideas

Words by themselves have connotations and **potential** denotations

- But only in the **act of uttering** a sentence do words get assigned specific meanings (referents)

This idea is attributed to the German logician **Frege**

- One of the founders of modern **logic**

4. Writing transforms discourse

Recall that **writing** and **speech**
are two modes of **discourse**

Writing **fixes** and **distances** discourse
(speech does not)

This produces **effects** on
each of the four traits

Effects of Writing on Discourse

	Speech (Dialog)	Effect of Writing
I	The event, the present, the fleeting moment	**The "meaning" of the event itself**, not the event, is captured. The locution, illocution, and perlocution to the extent that these can be **encoded** by **remappable** sentences.
II	Meaning and intention overlap	**Meaning is separated from intention**. The author becomes unimportant. The text stands on its own. Passive voice possible.
III	World and situation overlap Ostensive reference	**World is separated from situation**. Non-ostensive references form a structure. With text, “references open up the world”
IV	Addressed to a specific person or group	Addressed to **anyone who can read**. (Sometimes called "context collapse.")

These four effects constitute
the **objectivity** of the text

This makes it possible to be
both **scientific and interpretive**

Plurivocity

For all these reasons there is a problem of interpretation not so much because of the incommunicability of the psychic experience of the author, but because of the very nature of the verbal intention of the text. This intention is something other than the sum of the individual meanings of the individual sentences. A text is more than a linear succession of sentences. It is a cumulative, holistic process. This specific structure of the text cannot be derived from that of the sentence. Therefore the kind of "plurivocity" which belongs to texts as texts is something other than the polysemy of individual words in ordinary language and the ambiguity of individual sentences. This plurivocity is typical of the text considered as a whole, open to several readings and to several constructions.

Ricoeur describes the text as a **multidimensional cube** with many possible interpretations -- it is **plurivocal**

It can only be seen from **one perspective** at a time

In effect, he is describing **text as high-dimensional space** -- of sentences and relations between sentences (not just words)

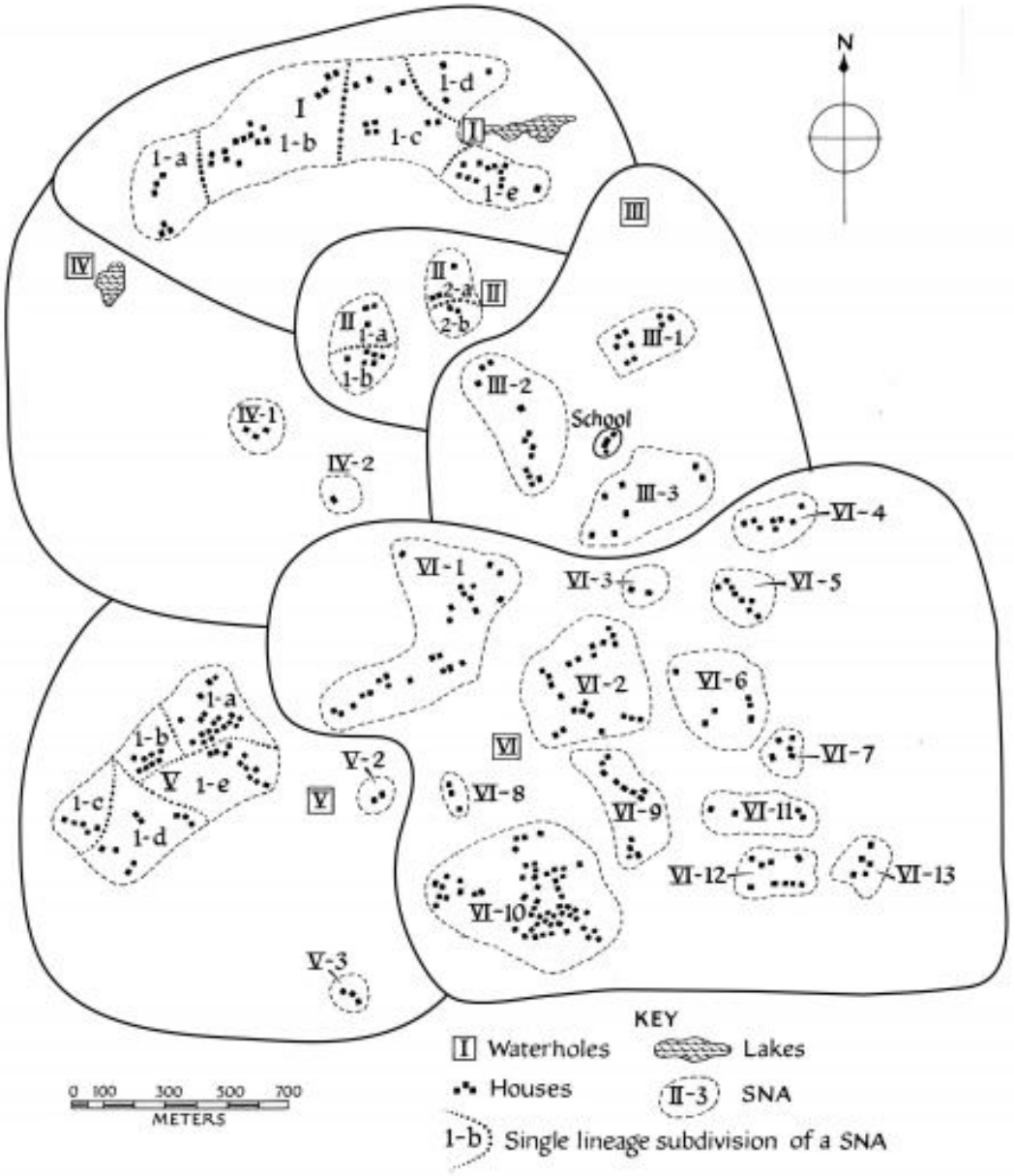
In other words, a text is a kind of **fossil**.

Just as archaeologists are able to reconstruct facts about an organism from its bones or a society from its **remains**, so can we reconstruct "worlds" from text.

And, indeed, Ricoeur argues that the archaeologist is performing a kind of hermeneutics.

Or, **a text is like a settlement pattern** in archaeology. The material **traces** of human activity form a structure that outlives the people who created it.

The **spatial and temporal distribution of these traces** can be used to infer things like social organization and population movements.

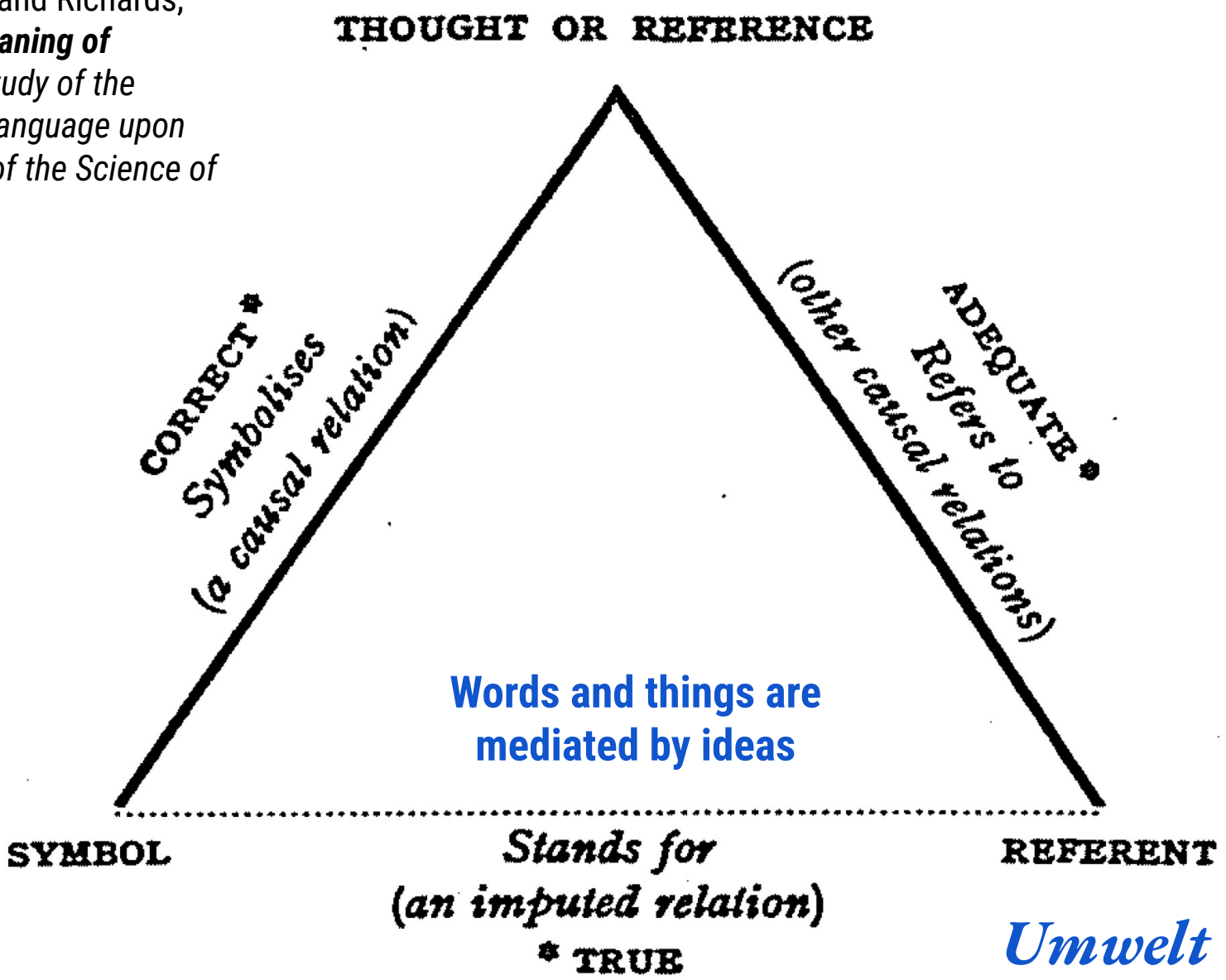


Most interesting is what happens to **reference**

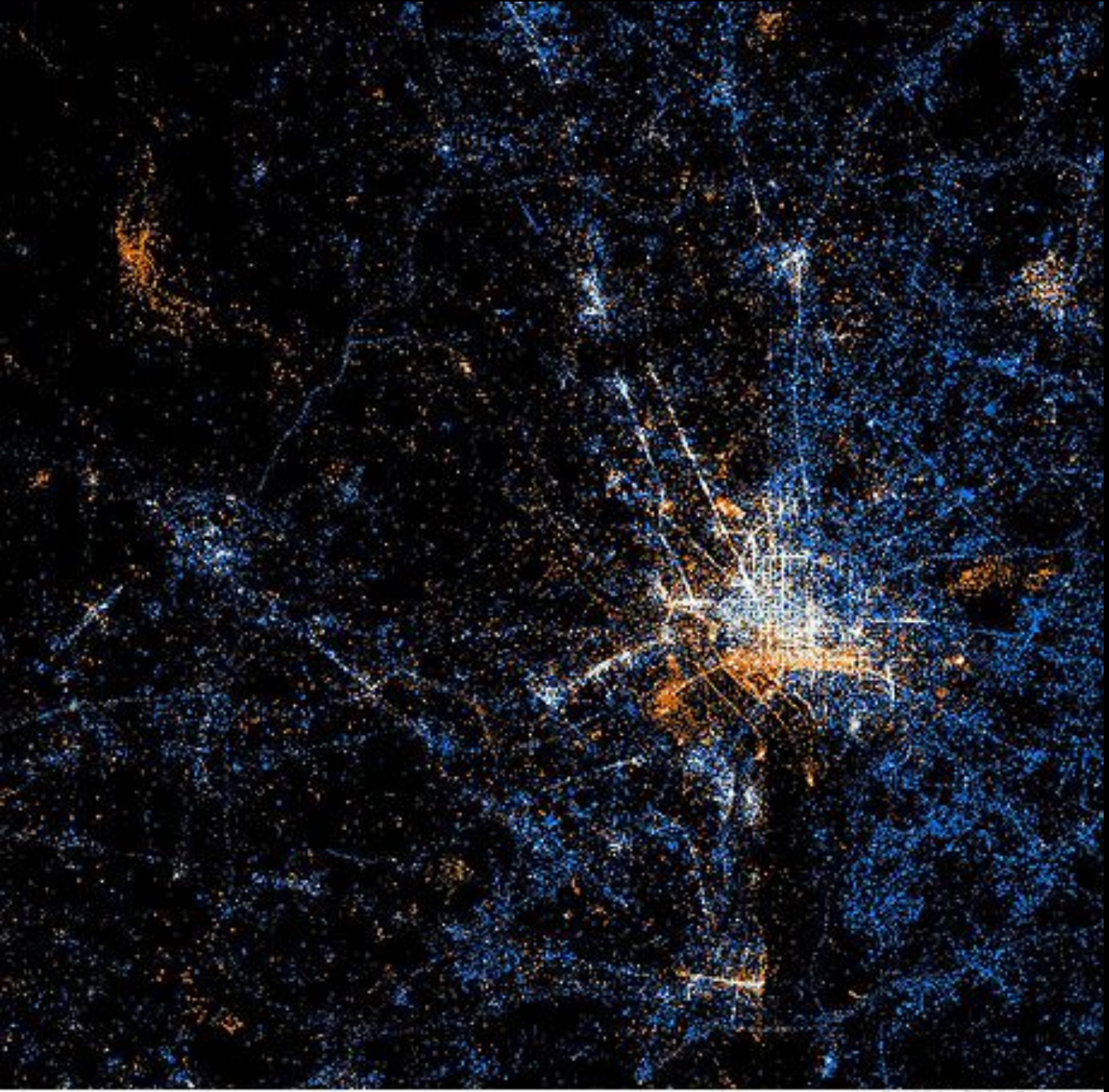
dimensions of our being-in-the-world. For me, this is the referent of all literature; no longer the *Umwelt* of the ostensive references of dialogue, but the *Welt* projected by the nonostensive references of every text that we have read, understood, and loved. To understand a text is at the same time to light up our own situation, or, if you will, to interpolate among the predicates of our situation all the significations which make a *Welt* of our *Umwelt*. It is this enlarging of the *Umwelt* into the *World* which permits us to speak of the references *opened up* by the text—it would be better to say that the references *open up* the world. Here again the spirituality of discourse manifests itself through writing, which frees us from the visibility and limitation of situations by opening up a world for us, that is, new dimensions of our being-in-the-world.

Welt

From: Ogden and Richards, 1923, ***The Meaning of Meaning***: *A Study of the Influence of Language upon Thought and of the Science of Symbolism*

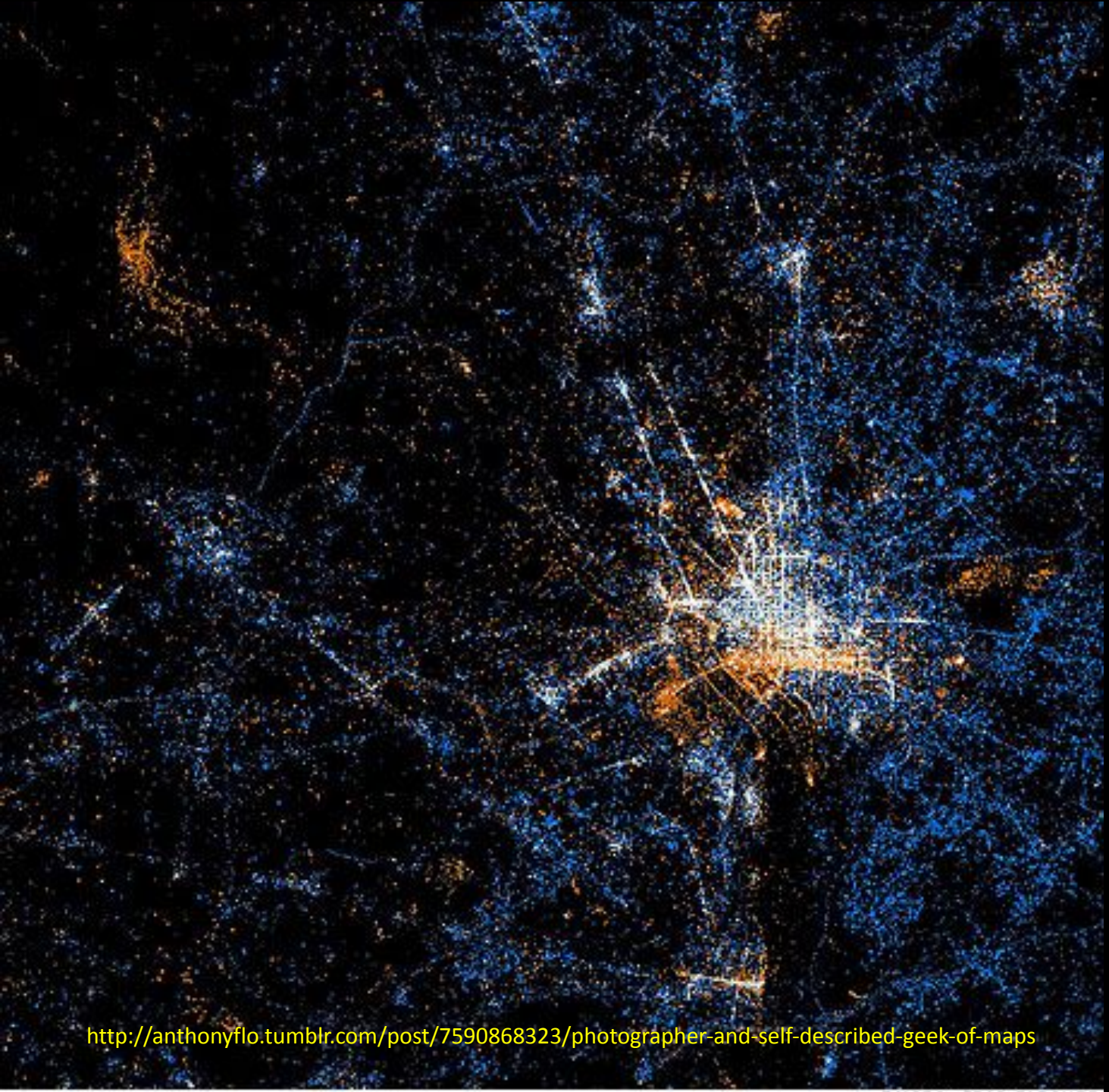


Umwelt



What is this an image of?

What do the points consist of?

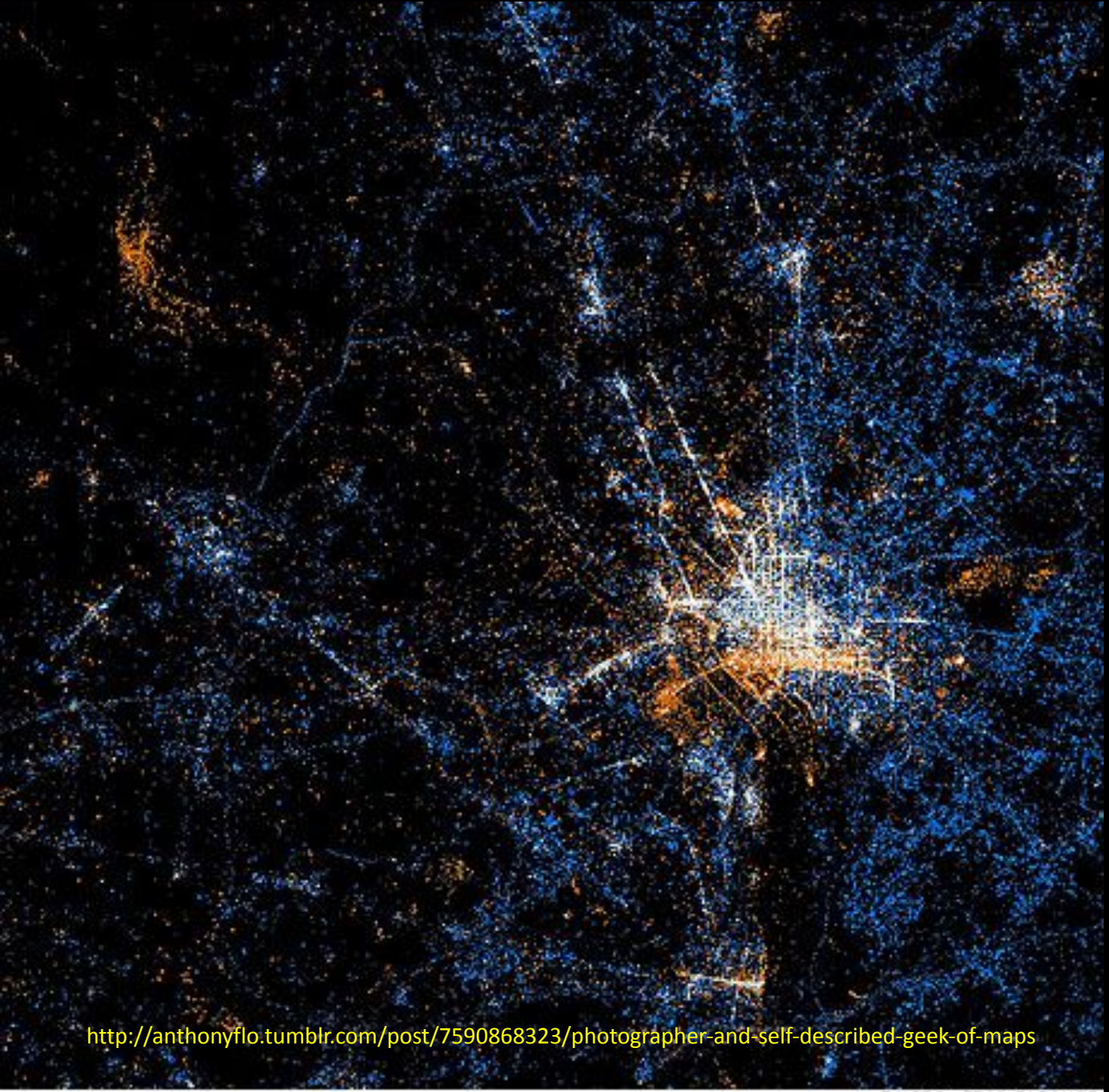


http://anthonyflo.tumblr.com/post/7590868323/photographer-and-self-described-geek-of-maps

This is a map of DC generated by thousands of individual Flickr and Twitter events.

The picture is a kind of signal—collective and unconscious, yet meaningful.

The patterns discerned from the signals are not intentional, but they are the products of intentional activity.



http://anthonyflo.tumblr.com/post/7590868323/photographer-and-self-described-geek-of-maps

This is something like a text.

It is the fixed remains of human behavior, captured and encoded as data, and interpretable as a structure.

It "refers" to a city, but it can also be the basis for a general theory of human settlement.

5. The structure of texts reflects the effects of fixation and distanciation

Effects of Fixation and Distantiation

Texts are materially encoded in **documents**

Documents can be **detached from the situation** of utterance

So, they take on **a life of their own**

Interpretation becomes a problem! Religious and legal texts ...

To **compensate** for the absence of shared situation ...

grammar develops, ex: can't use verbal tone, so hence punctuation etc.

Language **codes** get more **elaborate** — explicit **grammar** develops

Documents develop complex conventions and **structures**

Libraries form a new situation, as it were

Above all, **texts now require interpretation**

Back to Sentences

The grammatical structure of sentences reflect the structure of "**intentional exterioriziation**"

- The exteriorization of intention, objectification of thought, *noema*
- Subject, Verb, Object, etc. → Self, Other, World
- (Note we are using grammar here in the traditional sense)

This exteriorization makes **written communication** possible

- We can infer intentions, references, and effects from the structure of sentences to a degree

Text interpretation of texts is a **probabilistic exercise**

It moves from **guessing** to **validation**,
based on what is **known** about the **world** of the text
and the **grammar** of the language

Not all interpretations are valid --
a text is literally **objective**,
"a limited field of possible constructions"

This process is called the **hermeneutic circle**

Summary

Text and speech are both **discourse**, but they are not identical

- Text is **fixed** and **distanced** from the situation of the **speech event** and becomes an object unto itself
- Only what can be encoded **remains**

Text "contain" **worlds** that can be inferred through a circular process of **guessing** and **validation**

- This is not a process unique to text and text analytics
- In **archaeology**, the material remains of human behavior have this quality too

Text has a **structure** that can be described **formally . . .**

The OHCO Model of Text

What is a Text?

Mylonas and Renear, in effect, take up where Ricoeur leaves off

- Their goal is to describe **the essential qualities** of text as actually found in **documents**

In order to **represent text on the computer** for a variety of purposes

- Such as authoring, publishing, information retrieval ... and analytics
- Not lose information in the transfer

This task is **non-trivial**

- Many models have been tried and still exist
  e.g. PDF, plain text, XML, JSON, YAML

different ways of encoding text/data

A note about the readings

The two essays were written **just as the Web** was being invented

Some things sound archaic!

We take the results of these debates **for granted**

We all use HTML, for example, which adopts the OHCO model

By revisiting these debates we can better understand **the decisions we make** when we represent texts as data

Representational authenticity is essential to data science

Summary

1. Documents **fix** discourse as text
2. Texts are composed of **content objects**
3. Content objects are organized into **hierarchies**
4. **XML** implements the OHCO model
5. **Advantages** and **disadvantages** of OHCO
6. OHCO can be transformed into **table** form

1. Documents Fix Discourse as Text

The process of **fixing** discourse is **entextualization**

Documents are **physical**, texts are "**logical**"

- Texts can be extracted from documents and **re-entextualized**
- Leads to a form / matter distinction (**hylomorphism**)
- Texts are physical, but not *material* (next slide)

Texts have a **structure**

talking about text as information, so text is information and can be copied....

two editions of books, both same words are two different documents but two same texts bc it provides the same information

“Information is information, not matter or energy.”

— **Norbert Wiener, Cybernetics: or the Control and Communication in the Animal and the Machine**

tags: information

2. Texts are Composed of Content Objects

Content objects are **units of discourse**

- Paragraphs, sentences, chapters, etc.

Content objects are signified in documents by **visual codes**

- Indents, spacing, italics, font size, color, etc.

tone/sentiment of sentence

These codes capture some of the **illocutionary features** of the discourse event (described by Ricoeur)

- Grammar
- Looking ahead: we may want to capture these to do **sentiment analysis!**

E. C. Barksdale. *Daggers of the Mind: Structuralism and Neuropsychology in an Exploration of The Russian Literary Imagination.* Lawrence, Kansas: Coronado Press, 1979. 212 pp.

"Would it not be tantalizing," Barksdale asks, "to be present at the moment of the creation of a great work of literature?" It would indeed, and with this attempt to show how to do so, Barksdale shows himself to be a most ambitious literary theorist. His intent in this book is to do no less than establish a completely new method of literary study by joining Structuralism with the science of neuropsychology to form a new entity called Neurostructuralism. And more, to supplement traditional rhetorical criticism with another form of criticism he calls *poesis:* "the process of creativity" (9), "the study of the imagination" (22), "the study of human creativity as a product of the human mind and as an artifact of the human imagination" (18).

To this end, *Daggers of the Mind* is a two-fold study. Part one of the book, "Structures of the Mind," is an elucidation of the complexities of Structuralism and the even more complex essentials of neuropsychology. Barksdale attempts to synthesize here what he believes are the valid contributions of Structuralism with what has been discovered about the workings of the human brain and the human mind by neuropsychologists. Part two, "The Russian Imagination," is an application of Neurostructural theory to Barksdale's own area of specialization, nineteenth-century Russian literature. The first part deals with such subjects as "The Forgotten Mystery Solved: The Neurostructural Key," "Structuralism: The First Half of the Puzzle," "A Neurostructural Picture of the Mind," "States of the Mind," "Memory and the Mind," "Control from the Future: The Imagination," and "The Narrative Personality." The second part comprises Neurostructural criticism of six Russian writers: "Pushkin and the NonEncoded Engram," "Gogol: The Descent into Dreams," "Turgenev: Consciousness Regained," "Tolstoy: Consciousness Expanded," "Chekhov: Transcendence Lost," and "Dostoevsky:

THE STRUCTURAL STUDY OF MYTH

By Claude Lévi-Strauss

"It would seem that mythological worlds have been built up only to be shattered again, and that new worlds were built from the fragments."

Franz Boas, in Introduction to James Teit, *Traditions of the Thompson River Indians of British Columbia,* Memoirs of the American Folklore Society, VI (1898), 18.

1.0. Despite some recent attempts to renew them, it would seem that during the past twenty years anthropology has more and more turned away from studies in the field of religion. At the same time, and precisely because professional anthropologists' interest has withdrawn from primitive religion, all kinds of amateurs who claim to belong to other disciplines have seized this opportunity to move in, thereby turning into their private playground what we had left as a wasteland. Thus, the prospects for the scientific study of religion have been undermined in two ways.

1.1. The explanation for that situation lies to some extent in the fact that the anthropological study of religion was started by men like Tylor, Frazer, and Durkheim who were psychologically oriented, although not in a position to keep up with the progress of psychological research and theory. Therefore, their interpretations

Tables fussent mises a tire
En .ij. palés, la nel quer dire
Ains en tot .v. salles pleines
Si que l'en pooit a grant peines
Voie antre les tables avoir.
A chascune table por voir
Avoit ou roi, ou duc, ou conte.
.C. chevaliers tot par conte
En chascune table seoient.
Mil chevaliers de pein servoient
Et mil de vin et mil de mes
Vestuz d'ermins peliçons fres.
De mes divers toz sont serviz.
Et pour quant se ge nes vos di
Vos sauroie bien reison randre
Mes il mestuet a el antandre.

Explycyt li romans
D'Erec et d'Enyde.

Puis que ma dame
De Champaigne
Vialt que romans
A feire anpraigne,
Je l'anprendrai
Mult volentiers
Come cil qui
Est suens antiers
De quant qu'il puet el monde feire
Sanz rien de losange avant treire.
Mes tex s'an poïst antremetre
Qui i volsist losenge metre,
Si deïst, et je l'tesmoignasse,
Que ce est la dame qui passe
Totes celes qui sont vivanz,
Si con li funs passe les vanz
Qui vante en mai ou en avril.
Par foi, je ne sui mie cil
Qui vuelle losangier sa dame.
Dirai je tant com une jame
Vaut de pailes et de sardines,
Vaut la contesse de reïnes?
Naie voir, je n'en dirai rien,
S'est il voirs maleoit gré mien,
Mes tant dirai ge que mialz oevre
Ses comandemanz an ceste oevre
Sans ne painne que g'i mete.
Del chevalier de la charrete
Comance Crestiens son livre;
Matiere et san li done et livre

Kex qui ot servi as tables
Manjoit avoec les conestables.
La ou Kex seoit au mangier,
A tant ez vos .j. chevalier
Qui vint a cort mult acesmez,
De totes ses armes armez.
Li chevaliers a tel conroi
S'an vint jusque devant le roi
La ou antre ses barons sist.
Nel salua pas, einz li dist:
Rois Artus, j'ai en ma prison
De ta terre et de ta meison
Chevaliers, dames et puceles;
Mes ne t'an di pas les noveles
Por ce que jes te vuelle randre;
Einçois te voel dire et aprandre
Que tu n'as force ne avoir
Par quoi tu les puisses avoir.
Et saches bien qu'ainsi morras
Que ja aidier ne lor porras.
Li rois respont qu'il li estuet
Sofrir s'amander ne le puet,
Mes mult l'an poise durement.
Lors fet li chevaliers sanblant
Qu'aler s'an voelle, si s'an torne,
Devant le roi plus ne sejorne,
Et vient jusqu'a l'uis de la sale;
Mes les degrez mie n'avale,
Einçois s'areste et dit des la:
Rois, s'a ta cort chevalier a

2

Mr. Bennet was among the earliest of those who waited on[1] Mr. Bingley. He had always intended to visit him, though to the last always assuring his wife that he should not go; and till the evening after the visit was paid, she had no knowledge of it. It was then disclosed in the following manner. Observing his second daughter employed in trimming a hat,[2] he suddenly addressed her with,

"I hope Mr. Bingley will like it, Lizzy."

"We are not in a way to know *what* Mr. Bingley likes," said her mother resentfully, "since we are not to visit."

"But you forget, mama," said Elizabeth, "that we shall meet him at the assemblies,[3] and that Mrs. Long has promised to introduce him."

"I do not believe Mrs. Long will do any such thing. She has two nieces of her own.[4] She is a selfish, hypocritical woman, and I have no opinion[5] of her."

"No more have I," said Mr. Bennet; "and I am glad to find that you do not depend on her serving you."[6]

Mrs. Bennet deigned not to make any reply; but unable to contain herself,[7] began scolding one of her daughters.

"Don't keep coughing so, Kitty, for heaven's sake! Have a little compassion on my nerves. You tear them to pieces."

"Kitty has no discretion in her coughs," said her father; "she times them ill."

"I do not cough for my own amusement," replied Kitty fretfully.

"When is your next ball to be, Lizzy?"[8]

1 Visited, called on.

2 Clothing was extremely expensive in this period, so it was common for women to refurbish various items of apparel. Neil McKendrick points out that even hats and dresses preserved in museums show repeated signs of alteration. Fashions changed rapidly, another reason for efforts at home refashioning ("The Commercialization of Fashion," in *The Birth of a Consumer Society: The Commercialization of Eighteenth-Century England*, ed. Neil McKendrick, John Brewer, and J. H. Plumb [London: Europa, 1982], 43).

Retrimming hats in ways that corresponded to changing fashion or that altered color schemes or styles could make an old accessory seem new. Austen, at the age of twenty-two, writes to her sister, "next week [I] shall begin my operations on my hat, on which you know my principal hopes of happiness depend" (To Cassandra, 27 October 1798). A few months later, she specifies: "Flowers are very much worn, & Fruit is still more the thing.—Eliz: has a bunch of Strawberries, & I have seen Grapes, Cherries, Plumbs & Apricots—There are likewise Almonds & raisins, french plumbs & Tamarinds at the Grocers, but I have never seen any of them in hats" (To Cassandra, 2 June 1799). Nine days later: "We have been to the cheap Shop, & very cheap we found it, but there are only flowers made there, no fruit—& as I could get 4 or 5 very pretty sprigs of the former for the same money which would procure only one Orleans plumb . . . I cannot decide on the fruit till I hear from you again.—Besides, I cannot

Main Form

Phonetic Pronunciation

Parts of Speech formed from the word. "Sb.: means only substantive exists.

These numbers indicate 14th and 15th century variants

17th century variant forms or spellings

passant ('pæsənt), *a.* (*sb.*) Also 4–5 **-aunt, -e,** 7 **-ent.** [a. F. *passant*, pr. pple. of *passer* to PASS.]

The tiny cross indicates an archaic word. Next, the number "1" indicates this material applies to the *first* common meaning only. Other numbered meanings may follow.

†**1.** Surpassing, exceeding; excelling; = PASSING *ppl. a.* 3. *Obs.*

The status of the word appears here as "Obs." meaning "obsolete"

c **1386** CHAUCER *Knt.'s T.* 1249 Ffor euery wight that.. wolde his thankes han a passant name Hath preyd þat he myghte been of that game. **1413** *Pilgr. Sowle* (Caxton 1483) v. v. 76 The stones sholde nought haue kept them fro syngynge, for the passaunt ioye. *c* **1485** *Digby Myst.* v. 612 *Mynde.* Coryous aray I wyll euer haunt. *Vnderstondyng.* And I, ffal[s]nesse, to be passaunt.

Secondary meanings appear numbered sequentially with their own annotation

2. Passing, transitory, transient, fugitive. *Obs.*

c **1400** tr. *Secreta Secret., Gov. Lordsh.* 57 Coueyte noght þinges coruptibles & passant. **1604** WEBSTER *Ode* in *Arch's of Triumph*, For pleasure's stream Is like a dream, Passant and fleet, as is a shade. *a* **1677** BARROW *Wks.* (1686) II. Serm. xvi. 223 Our actions (even our passant words, and our secret thoughts). **1715** JANE BARKER *Exilius* II. II. 55 All the Glories of this World are passant.

The "c." (Latin circe) indicates an approximate date. The numbers in bold print refer to the first recorded date the OED editors have found for the word's appearance in written form with this particular meaning. If the work comes from a famous author used as a standard dating marker, that author's name appears in all capital letters, along with an abbreviated version of the word's title and line numbers or pages (if applicable). These entries are listed chronologically under each individual meaning. Entries for a different meaning have a separate list with their own chronological entries -- such as in entry #2 here.

Letter 1

To Mrs. Saville, England.

St. Petersburgh, Dec. 11th, 17—.

You will rejoice to hear that no disaster has accompanied the commencement of an enterprise which you have regarded with such evil forebodings. I arrived here yesterday, and my first task is to assure my dear sister of my welfare and increasing confidence in the success of my undertaking.

I am already far north of London, and as I walk in the streets of Petersburgh, I feel a cold northern breeze play upon my cheeks, which braces my nerves and fills me with delight. Do you understand this feeling? This breeze, which has travelled from the regions towards which I am advancing, gives me a foretaste of those icy climes. Inspirited by this wind of promise, my daydreams become more fervent and vivid. I try in vain to be persuaded that the pole is the seat of frost and desolation; it ever presents itself to my imagination as the region of beauty and delight. There, Margaret, the sun is for ever visible, its broad disk just skirting the horizon and diffusing a perpetual splendour. There—for with your leave, my sister, I will put some trust in preceding navigators—there snow and frost are banished; and, sailing over a calm sea, we may be wafted to a land surpassing in wonders and in beauty every region hitherto discovered on the habitable globe. Its productions and features may be without example, as the phenomena of the heavenly bodies undoubtedly are in those undiscovered solitudes. What may not be expected in a country of eternal light? I may there discover the wondrous power which attracts the needle and may regulate a thousand celestial observations that require only this voyage to render their seeming eccentricities consistent for ever. I shall satiate my ardent curiosity with the sight of a part of the world never before visited, and may tread a land never before imprinted by the foot of man. These are my enticements, and they are sufficient to conquer all fear of danger or death and to induce me to commence this laborious voyage with the joy a child feels when he embarks in a little boat, with his holiday mates, on an expedition of discovery up his native river. But supposing all these conjectures to be false, you cannot contest the inestimable benefit which I shall confer on all mankind, to the last generation, by discovering a passage near the pole to those countries, to reach which at present so many months are requisite; or by ascertaining the secret of the magnet, which, if at all possible, can only be effected by an undertaking such as mine.

3. Content Objects are Organized into Hierarchies

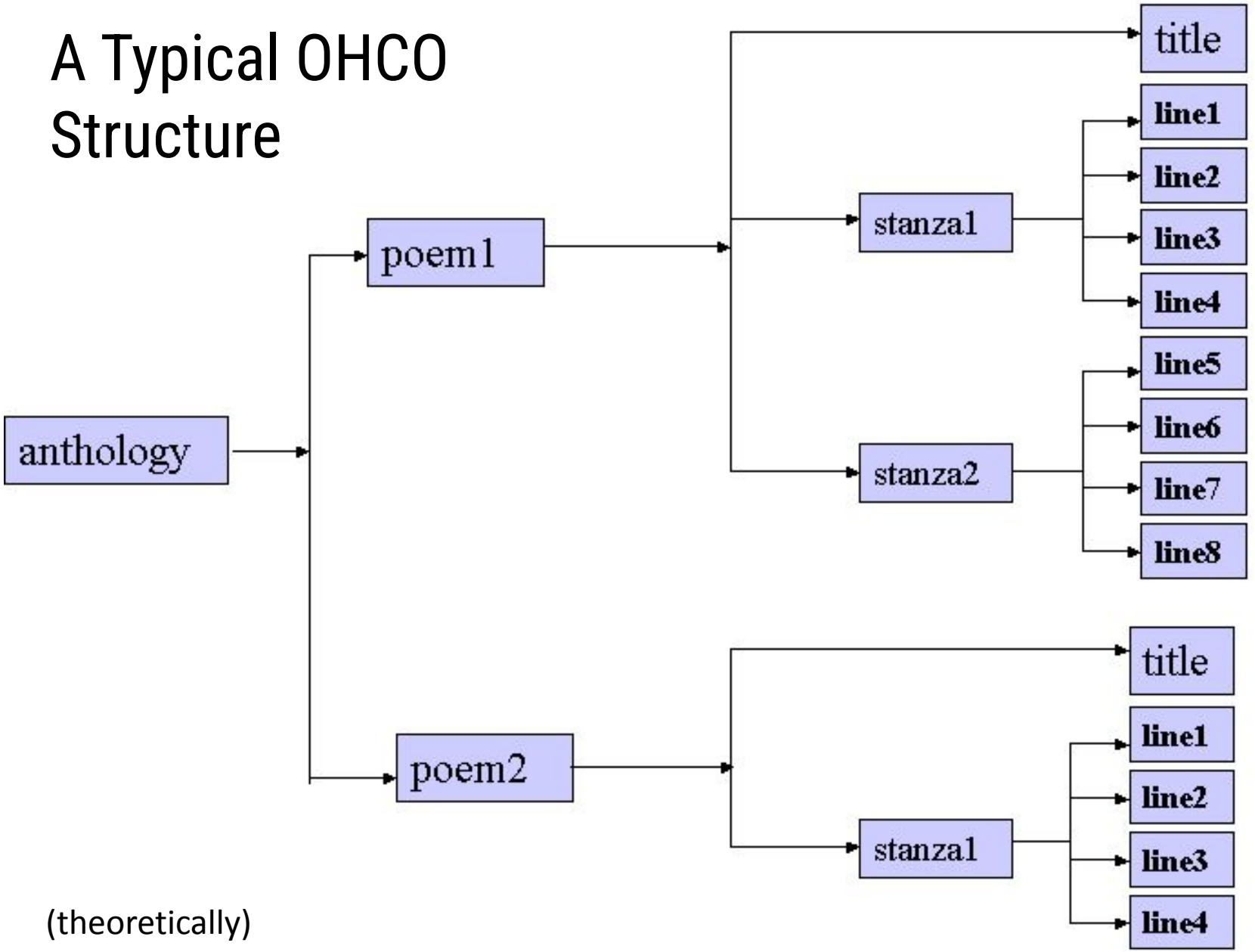
Content objects in a text compose a **tree-like structure**

- Content objects are like **containers** that can contain other containers

Texts can classified by the **structure** of this tree

- “Document types” differ in the kinds of content types and how they can be arranged
- This is called an **ordered hierarchy**
- Hence **OHCO**, “ordered hierarchy of content objects”

A Typical OHCO Structure



(theoretically)

Other document types

- Play
  - Act +
    - Scene +
      - Line +
- Book
  - Chapter +
    - Verse +
- Letter
  - Heading
    - Return Address
    - Date
    - Recipient Info
      - Name
      - Title
      - Address
  - Content
    - Salutation
    - Paragraph +
    - Closing

4. XML Implements the OHCO Model

A goal of the OHCO model is to translate visual codes and hierarchical structures an **explicit language**

- Both human and machine readable

tagged based system

XML (and its predecessor SGML) is the most established implementation of OHCO

- It uses **tags** to explicitly represent content objects and their functions
- It also uses schema to specify the set of and structure of content objects

XML represents OHCO like this

```
<anthology>
  <poem><title>The SICK ROSE</title>
    <stanza>
      <line>O Rose thou art sick.</line>
      <line>The invisible worm,</line>
      <line>That flies in the night</line>
      <line>In the howling storm:</line>
    </stanza>
    <stanza>
      <line>Has found out thy bed</line>
      <line>Of crimson joy:</line>
      <line>And his dark secret love</line>
      <line>Does thy life destroy.</line>
    </stanza>
  </poem>
            <!-- more poems go here    -->
</anthology>
```

tag and element; tag is 'line' and element is the content

Notice how the element names reference units, not layout or style

```
/ID[<38D7748772979E3894AE57BB32451443><0F8D9F7F3BF04D95CAFC9B9B8E0AE5F8>]
/Info 38 0 R
/Prev 4319445
>>
startxref
0
%%EOF
44 0 obj
<</Type/Catalog/Pages 35 0 R/Metadata 41 0 R/PageLabels 39 0 R>>
endobj
78 0 obj
<</S 388/L 476/Filter/FlateDecode/Length 79 0 R>>
stream
xúc```b``©c`e` ºÀ ΔÄb@16  é
L
 @ Δoimù C ∞â°ù>√êπêsŒ 6 ô • ò÷4çYZÂıiœR Ö#")Çû¶¢ëÂ ªúx › E& ÜNo"∏èê*ìÃ‹Æÿe~ê√êUí%»◊
ïÎ '±∏F†Y pÇ´ ï—‘ ©ÀÓ†hØHÍ˜Ä09q⁄Ï" ı q î:¿@G@0ÿÄ@Å¡ÒÙi ≠ƒ*` g ^6áÕá/.åhç=Æ¥çß ª≈9á{ítÀ Ò ° ,
(ê¯a°†?Yas3œu∏˘’ ÔÁ2Ä„ò¡ é¸p÷
endstream
endobj
79 0 obj
268
endobj
45 0 obj
<<
/Type/Page
/Resources <</ProcSet[/PDF/Text/ImageB]/Font<</F1 46 0 R/F2 49 0 R/F3 52 0 R/F4 55 0 R/F
/MediaBox[0 0 612 792]
```

This is in contrast to **procedural** markup like PDF

5. Advantages and disadvantages of OHCO

Every element has **a precise address** in the text

E.g. `/html/body/p[1]`

Texts can be described in the language of **kinship**

Ancestors, parents, siblings, children, etc.

Texts can be restructured and manipulated by **known algorithms**

Traversing, Pruning, Cross-referencing

However, **not all aspects** of a text can be modeled as hierarchies

Relationship between physical and logical structure

Precisely such had the paragraph originally stood from the printer's hands; but Sir Walter had improved it by adding, for the information of himself and his family, these words, after the date of Mary's birth:—"Married, December 16, 1810, Charles, son and heir of Charles Musgrove, Esq. of Uppercross, in the county of Somerset," and by inserting most accurately the day of the month on which he had lost his wife.

Then followed the history and rise of the ancient and respectable family in the usual terms; how it had been first settled in Cheshire, how mentioned in Dugdale, serving the office of high sheriff, representing a borough in three successive parliaments, exertions of loyalty, and dignity of baronet, in the first year of Charles II. with all the Marys and Elizabeths they had married; forming altogether two handsome duodecimo pages, and concluding with the arms and motto:—"Principal seat, Kellynch Hall, in the county of Somerset," and Sir Walter's handwriting again in this finale:—

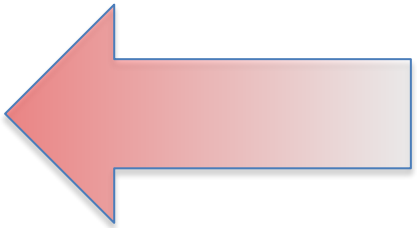
"Heir presumptive, William Walter Elliot, Esq., great grandson of the second Sir Walter."

Vanity was the beginning and end of Sir Walter Elliot's character: vanity of person and of situation. He had been remarkably handsome in his youth, and at fifty-four was still a very fine man. Few women could think more of their personal appearance than he did, nor could the valet of any new made lord be more delighted with the place he held in society. He considered the blessing of beauty as inferior only to the blessing of a baronetcy; and the Sir Walter Elliot, who united these gifts, was the constant object of his warmest respect and devotion.

His good looks and his rank had one fair claim on his attachment, since to them he must have owed a wife

Pages and Paragraphs

Physical and **logical** structures that **overlap**



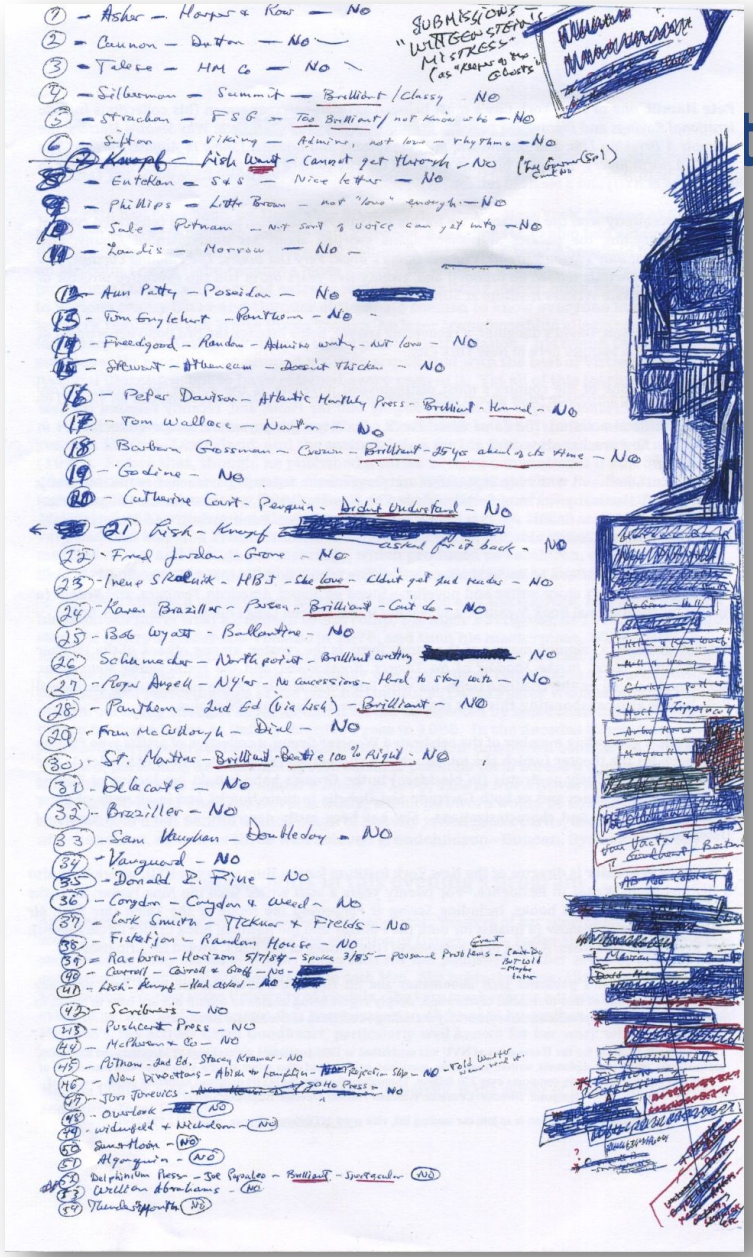
Solution 1: Split Elements

```
<page n="2">
...
<p id="foo">His good looks and his rank had one fair claim on his attachment, since to them he must have owed a wife</p>
</page>
<page n="3">
<p id="bar" prev_id="foo"> a very superior character to anything deserved by his own.</p>
...
</page>
```

Solution 2: Use “Milestones”

His good looks and his rank had one fair claim on his attachment, since to them he must have owed a wife <pb n="3" /> a very superior character to anything deserved by his own.

One structure gets backgrounded



II. In addition to the time he spent working with Jefferson, with Adams, almost daily, he

-30-

all four of
But for/Abigail and John, indeed for the presence of , as much as anything,
Adams family, it was Jefferson that made
the time in France one of the brightest,
happiest
interludes of their lives. He dined with them at Auteuil repeat-
edly, almost from the day the family
moved in, and they in turn with him, in Paris,
of his own
once he settled in a house in what was called the Cul de Sac
near the Opera.
Taitbout, off the Boulevard des Italiens,
kind of
John Quincy came to regard Jefferson's house as
refuge when in the city and struck up friendships with
David
Jefferson's two young aides, Colonel Humphreys
showed
and William Short, And the interest Jefferson
appreciated
Quincy, was not lost on his
parents. "He appeared to me as much your boy as mine," Adams
remind Jefferson years
afterward.

"Spe nt the evening with Mr.
John Quincy would record
Jefferson, whom I love to be w ith,"
in his diary, "because he is a man of very extensive
learning, and pleasing manners."

"Dined at Mr. Jefferson's, became a familiar entry in John
Quincy's diary. "In afternoon, w ent ... to Paris, Mr.
Jefferson's "Spent evening with Mr. Jefferson, whom I love
to be with, because he is a man of very extensive learning,
and pleasing manners."

"Came home and found Mr. Jefferson again," Nabby recorded
in her journal
another day. "He is an agreeable man."

A manuscript page from his upcoming book.

What about this?

The problem of overlap suggests that OHCO is **not as simple** as it may appear

However, many of these problem are **edge cases**

OHCO works well enough if you **choose** between physical and logical representations of text

6. OHCO can be transformed into table form

Another way to implement OHCO is with **dataframes** (tables)

This is the approach we will take in this class

This builds on a point made by Mylonas and Renear

Text analytics is a “value-added” function

INFORMATION RETRIEVAL FUNCTIONS: **The OHCO model treats documents and related files as a database of text elements that can be systematically manipulated. This can facilitate not only personal information retrieval functions, such as the generation of alternative views, but also a variety of ‘‘value-added’’ data retrieval functions.**

```
chap 0
    para 0
        sent 0
            token 0
            token 1
            token 2
            ...
        sent 1
            token 0
            token 1
            ...
        ...
    para 1
        ...
    para 2
        ...
    ...
chap 1
    ...
```

chap_num	para_num	sent_num	token_num	token_str
0	0	0	0	ETYMOLOGY
0	0	1	0	
0	0	1	1	.
0	0	1	2	
0	0	2	0	
0	1	0	0	
0	1	0	1	(
0	1	0	2	Supplied
0	1	0	3	
0	1	0	4	by
0	1	0	5	
0	1	0	6	a
0	1	0	7	
0	1	0	8	Late
0	1	0	9	
0	1	0	10	Consumptive
0	1	0	11	
0	1	0	12	Usher
0	1	0	13	

Markup Languages

Markup Languages

Markup languages are designed to **explicitly encode features** of text that humans can infer

They provide a means to format a text so that it may be **computationally processed** for a variety of purposes

- Rendering in different formats, e.g. HTML or PDF
- Combining, unbundling, etc., e.g. for document summarization
- Extraction of parts for indexing, querying, analytics

Markup Languages

Our interest in markup languages is **two-fold**:

- To learn about this fundamental form of text-encoding so that we can **process** it later
- To read the specifications as a **theory** of text

As a fundamental form of text-encoding

- XML and TEI represented **the most sophisticated means available** to represent text authentically

As a theory of text

- XML and TEI represent **a lot of thinking** on what constitutes a text

What is XML?

Stands for **eXtensible Markup Language**

- Actually invented after the web
- A simplification of SGML, the language used to create HTML

It specifies a set of rules for **creating specialized markup languages** such as HTML and TEI

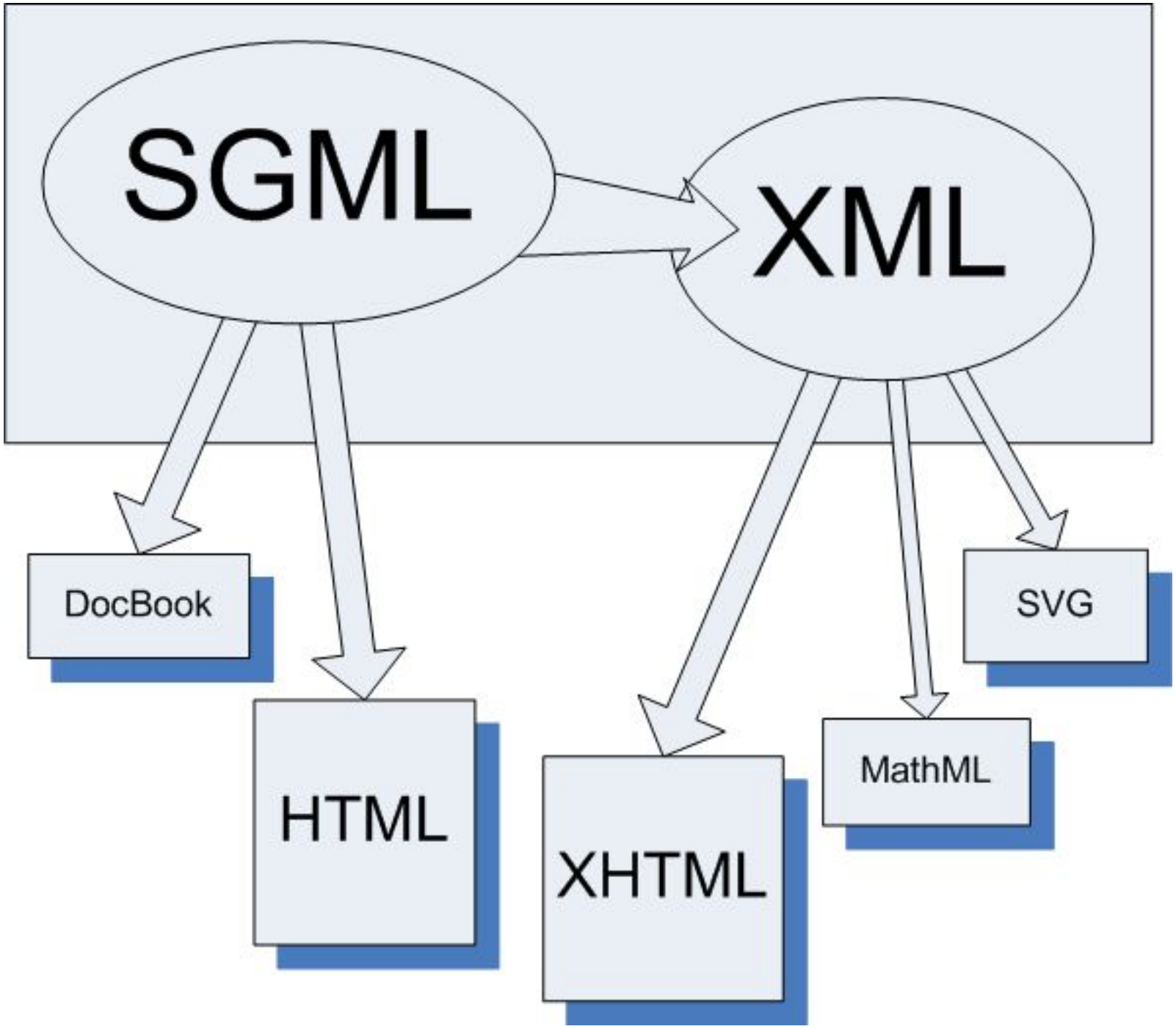
- It is simplified version of the SGML
- Standard Generalized Markup Language

SGML was invented in the early 1970s to wrest the control of documents from computer people who were taking over industries like law and accounting

XML

Many languages are written in XML

- XHTML
- SVG
- TEI
- MathML
- There's even Predictive Model Markup Language


SGML
XML
DocBook
HTML
XHTML
MathML
SVG

What is TEI?

TEI stands for **Text Encoding Initiative**

- Founded in **1987**

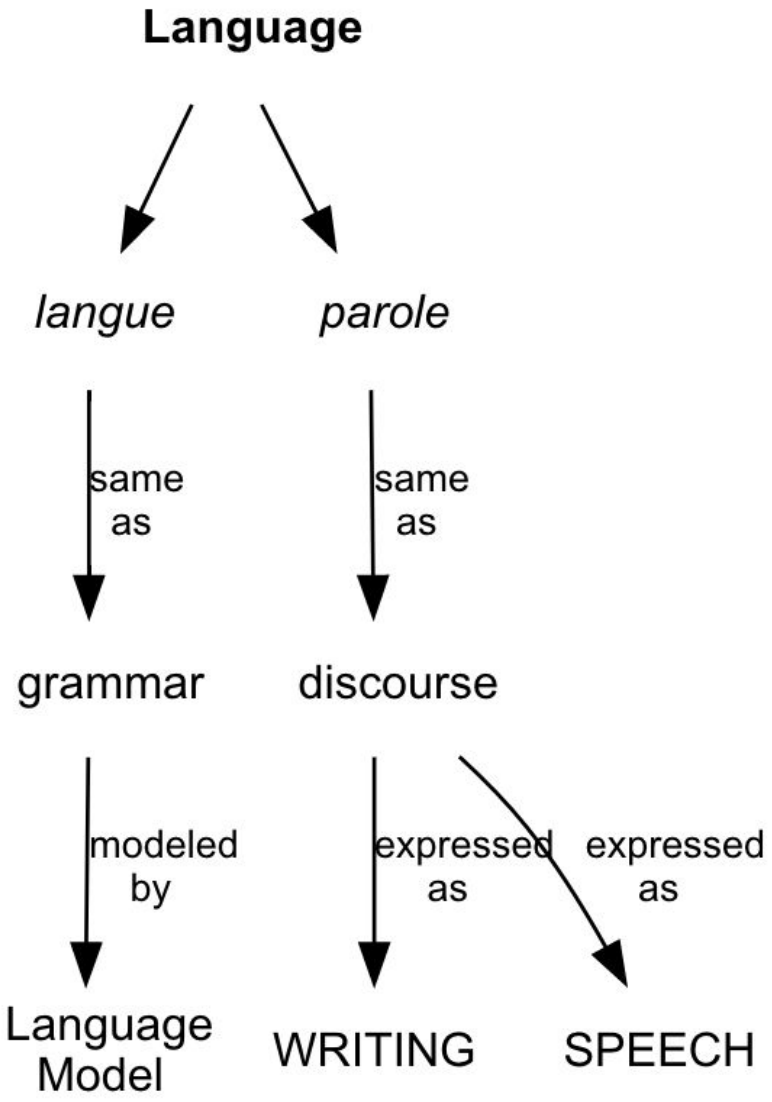
TEI is a markup language written in **XML**

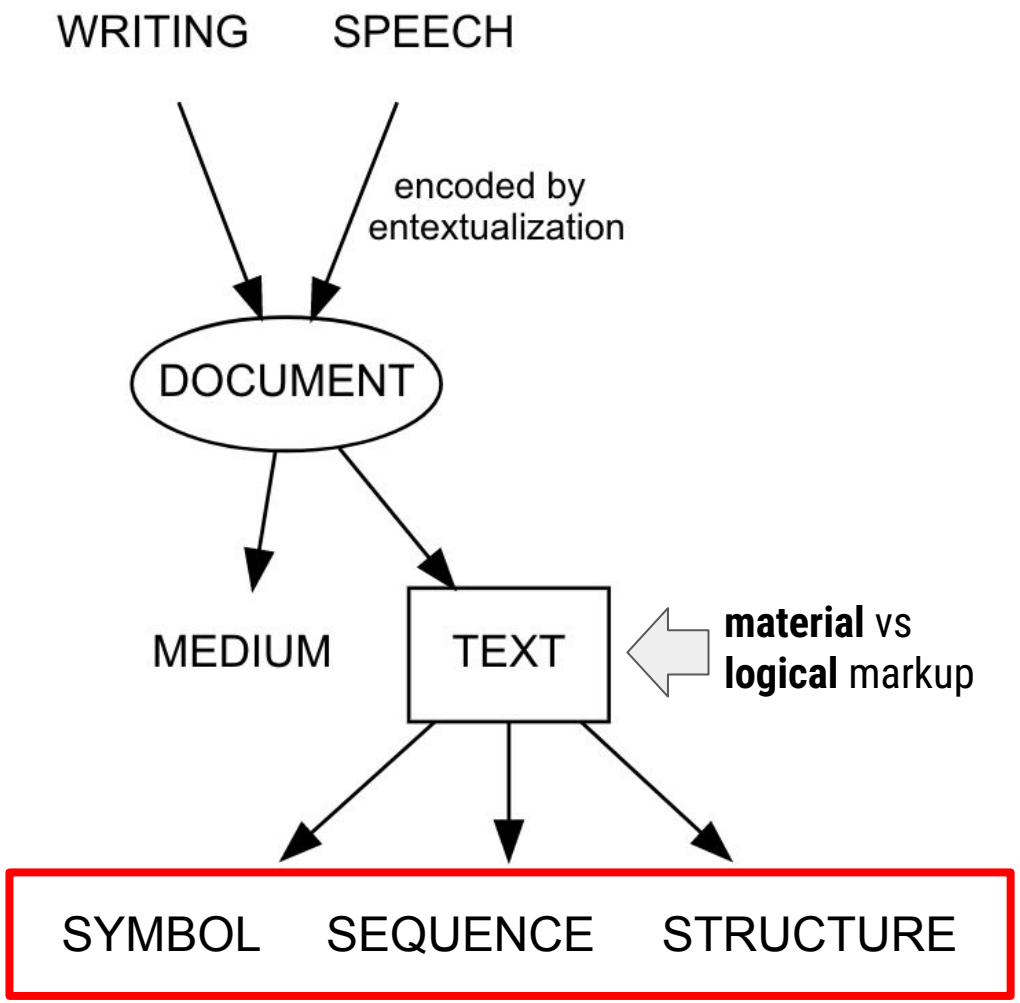
- First written in **SGML**
- Designed to encode **machine-readable texts** of interest to the **humanities and social sciences**

There are many text archives written in TEI

- UVA has developed many of them!

Summary


Language
langue
parole
same as
same as
grammar
discourse
modeled by
expressed as
expressed as
Language Model
WRITING
SPEECH


WRITING
SPEECH
encoded by entextualization
DOCUMENT
MEDIUM
TEXT
material vs logical markup
SYMBOL
SEQUENCE
STRUCTURE

Text Models