Vagueness - the Neglected Feature in Big Data

Walther von Hahn

Universität Hamburg • Computer Science Department

E-Mail: vhahn@informatik.uni-hamburg.de

Colloque international • Les mégadonnées et les sciences sociales. 22. – 24. Sept. 2017 • Bucarest

Contents

- Theoretical Background
- Which Big Data?
- Vagueness
  - types
  - on several layers
  - historical material
- Annotatations and Inferencing
- Data collection and interpretation in DH
- Vagueness in Interpretation GUIs
- **Lexical and Syntactic Sources of vagueness in original**
- **Example for necessary Manual Annotation of Factual uncertainity**
- Summary
- References

Overview

In social science (as in other fields of „Digital Humanities“) big data projects tend to collect data as facts in a (relational) data base. Social science, however partakes - as a humanity - according to Wilhelm Dilthey in a hermeneutic paradigm for establishing social hypotheses. Accordingly, social data often consist either of texts mirroring attitudes, allegations, beliefs, etc., or are reactions of test subjects to verbal stimuli. Such material cannot be treated as facts like numbers or positive propositions. On the other hand, analysing only formal features in the material does rarely contribute to the hermeneutic aims of the sociological quest.

-
- The talk is about possible ways out of this dilemma. A first solution is the subsequent usage of big data for human reading and interpretation only, which, however, underestimates the scientific power of computing.
-
- Another solution is a semi-automatic annotation of vagueness. It can be achieved by metadata about the credibility of texts and authors as well as by lexical annotations of vagueness expressions. Occurrences of “perhaps”, “mostly” or “to a certain extent” (to name only obvious examples) may support a fair social interpretation. Moreover, annotation supports semantic qualifications and allows for reasoning over vague features in big data.

Empirical Social Science, Humanity or Science

SCIENCE:

- statistics and stochastics
- computerized methods for (semi-automatic) collection, retrieval, annotation and analysis, data exchange, linking among data

HUMANITY:

- quantitative methods support humanities‘ qualitative research, but do not replace them. The main hermeneutic task is left open.
- Computer formalisms typically model **facts** in **data bases**. However, only few humanistic issues are facts, most are open to interpretation. Standard data bases in some way obscure the data by alleging “facts”. The main hermeneutic task is still left open.
- Example: Most owners of TV sets have a low IQ ➔ Background: Both facts are co-occurrent, who decides, that they are not causal

Science and Humanities

Wilhelm Dilthey (*Einleitung in die Geisteswissenschaft* 1922):

Dilthey describes history as “a series of world views.” Man can only understand himself through what “history can tell him” … never in objective concepts. Dilthey emphasizes the “intrinsic temporality of all understanding” i.e., that man’s understanding is dependent on past world views, interpretations, and a shared world.

Later on Hans-Georg Gadamer (*Wahrheit und Methode* 1960) declared, that interpreting a text involves a fusion of horizons (*Horizontverschmelzung*). Both the text and the interpreter find themselves within a particular historical tradition, or “horizon”. Each horizon is expressed through the medium of language, and both text and interpreter belong to and participate in history and language.

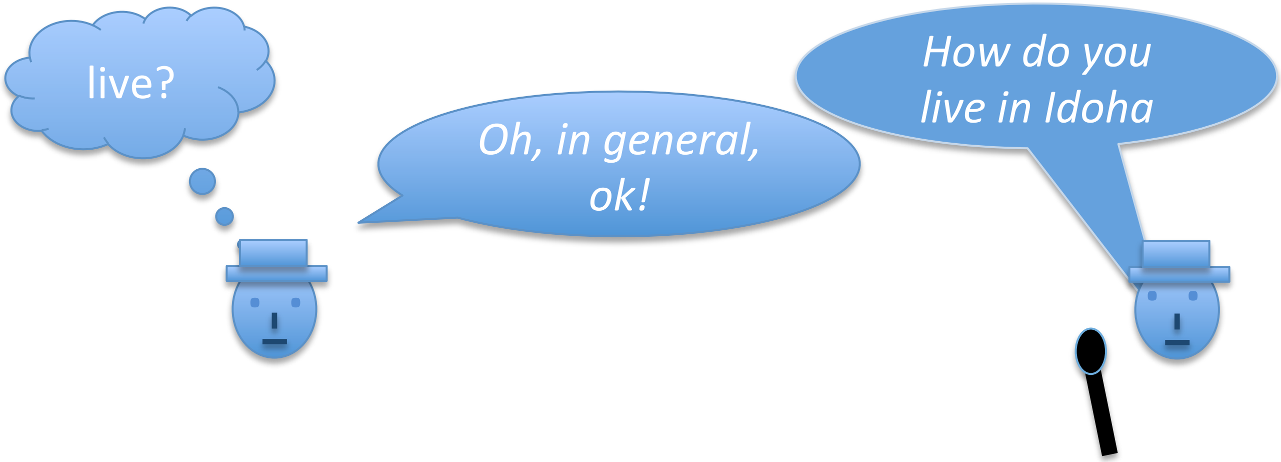
Jürgen Habermas (*Technik und Wissenschaft als Ideologie, Theorie des kommunikativen Handelns*, 1968) distinguishes between purposive rational action and social action, the latter being the proper subject of humanities.

Jürgen Habermas’ concept and theory of communicative rationality distinguishes itself from the rationalist tradition, by locating rationality in structures of interpersonal linguistic communication rather than in the structure of the cosmos.

Which Big Data?

- Vagueness in social science is an issue for those big data, which in the end are evaluated semantically, i.e. by analyses or annotations higher than linguistic formal structures.
- At least when you use wordNet synsets or even worse, their translations from English, you have to envisage vagueness problems, because you use word senses in a hermeneutic way, not only by measuring.

Dialogue and Semantic Interpretation



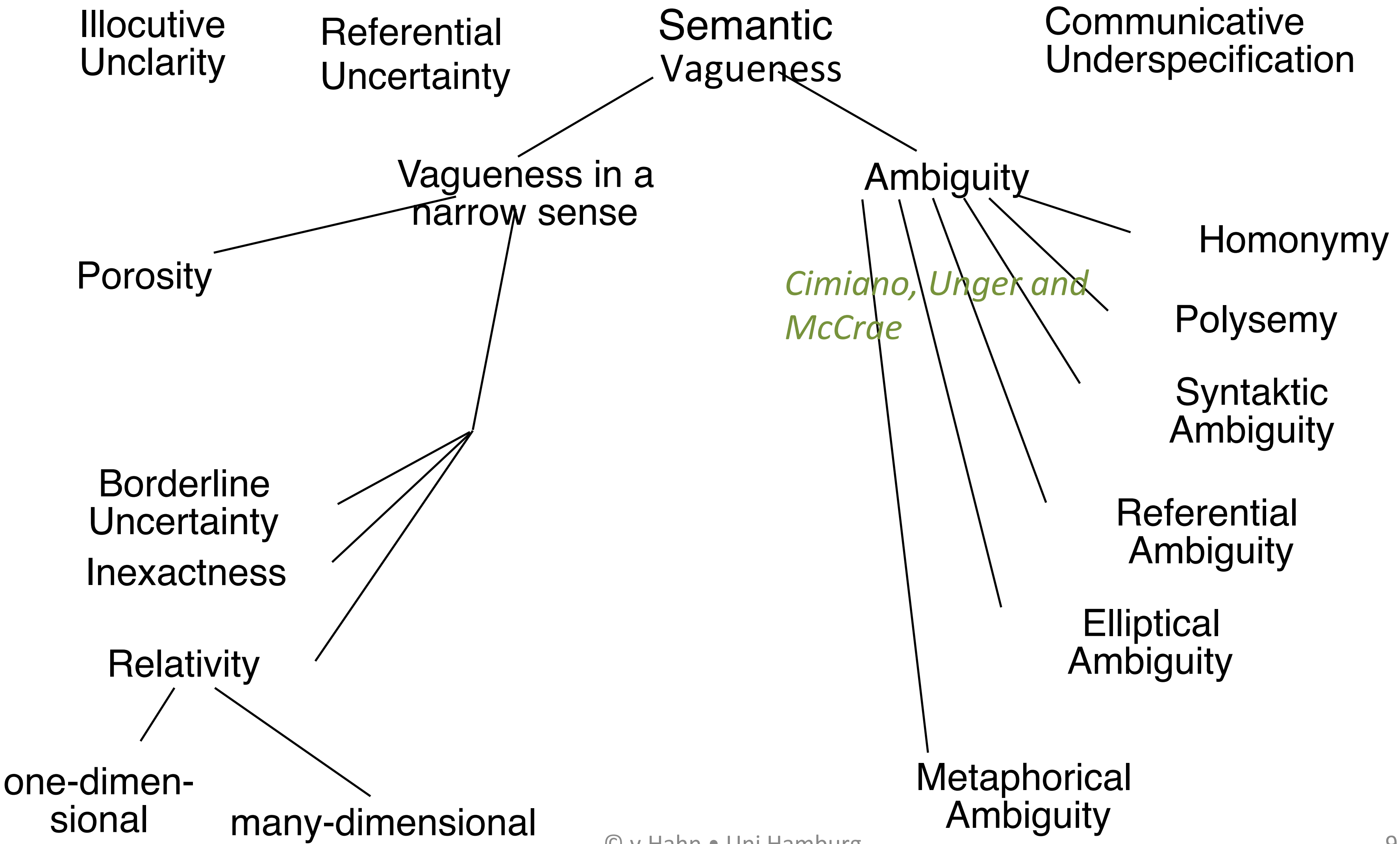
He is satisfied

this a fact in the data base

Example

- Many social science data come from public opinion polls and are individual responses to verbal stimuli.
- Measuring formal details (e.g. sentence length, response time) is not a hermeneutic activity.
- However, a later attachment of meaning (*„A majority of respondents are sceptic against African immigrants“*) to numerical results is a hermeneutic issue and subject to interpretation.
- To avoid invalid interpretations, user have to include into their evaluation metadata about the survey details, i.e. about questionnaire and the respondents.
- A linguistic analysis has to check the homogeneity of the random sample or possible interference among the interviewer and the respondents.
- Within big data a user has to merge the metadata into a data reliability info.

M.Pinkal‘s Schema of Semantic Vagueness



Illocutive Unclarity

Referential Uncertainty

Semantic Vagueness

Communicative Underspecification

Vagueness in a narrow sense

Ambiguity

Porosity

Borderline Uncertainty

Inexactness

Relativity

one-dimen-sional

many-dimensional

Homonymy

Polysemy

Syntaktic Ambiguity

Referential Ambiguity

Elliptical Ambiguity

Metaphorical Ambiguity

Cimiano, Unger and McCrae

Factual Uncertainty

(yet) unexplored facts	*"the moon is 384402,56 m distant from the earth”*
range expressions	*“The beginning of the 18. century” “Romania in the middle ages“*
uncertain definition	*“the northern slope of the mountain”*
Inexact measures	*„4 Tagereisen“, „10 Fuß“, a 4 days’ journey, 10 feet”*
unclear place	*„Syrfia“*
unclear facts	*„auf Befehl des Sultans“, „by order of the sultan“*
unclear time	*„In grauer Vorzeit“, „in prehistoric times“*
unclear person	*„der damalige Fürst“, „the former prince“*
unclear action	*„Die Unterwerfung der Barbaren“,* *„the submission of the barbarians“*

Challenge for DH: Vagueness on several layers

Examples from historic texts:

- Linguistic vagueness,
- Logical vagueness,
- Fuzzy concepts
  - “Before Stephan the Great, *all mountains around* Moldavia belonged to Transilvania and the country was *narrow* on this side”…,
- Vague or concurrent ontologies:
  - The Turkish and the Moldavian administration,
- Referential vagueness or uncertainty
  - The origin of the hill “Chan Tepesi” or “Mogila Rabuy”,
- Naïve History (derived from ‘naive physics’)
  - „The Roman Empire conquered Dacia“,
- Historical change,
- Vagueness of the sources

The more you go into history, the more data become vague

- measures
- time span expressions: *In the beginning of Xth century, shortly later*
- Persons: *the former prince, the current pope*
- even NEs are often vague,

are vague,

- Additionally, changes in writing creates artificial vagueness,

How to annotate

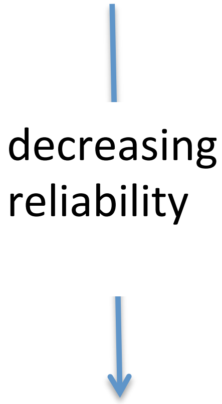
- In big data you cannot annotate large amounts of texts with reasonable costs.
- The only ways out:
  - small learning texts and automatic propagation,
  - automatic annotation of lexical indicators,
  - including meta-data for text classes,
  - establishing inference rules for „vagueness combinations“ .

Lexical Vagueness Predictors

- Modal verbs: must, should, will, can ...
- Adverbs: *perhaps, for example, so to say, possibly, maybe, by any chance, roughly, rude, coarse, and so on, and so forth, basicly, ...*
- Adjectives: *simplified,*
- Comparative degrees: *better, more, worse...*
- Vague quantifiers: *many, most, mostly, majority, often*

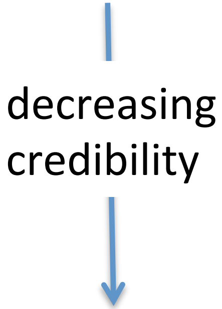
Metadata to be included in the GUI

genre:
- official document,
- letter
- fiction,
- fairy tale
- legend,
- folk tradition



decreasing reliability

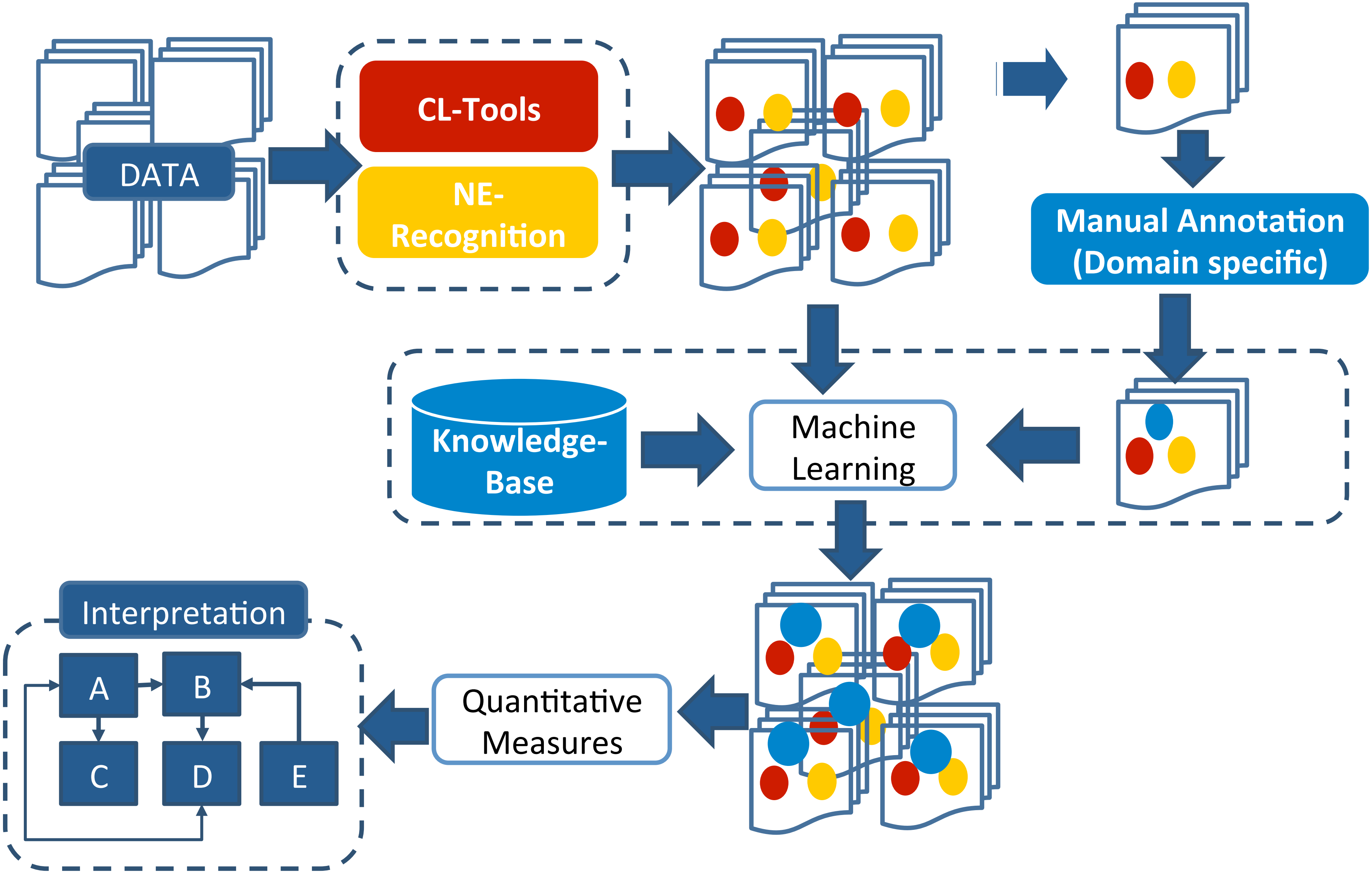
credibility of author
- politician
- journalist
- fiction writer



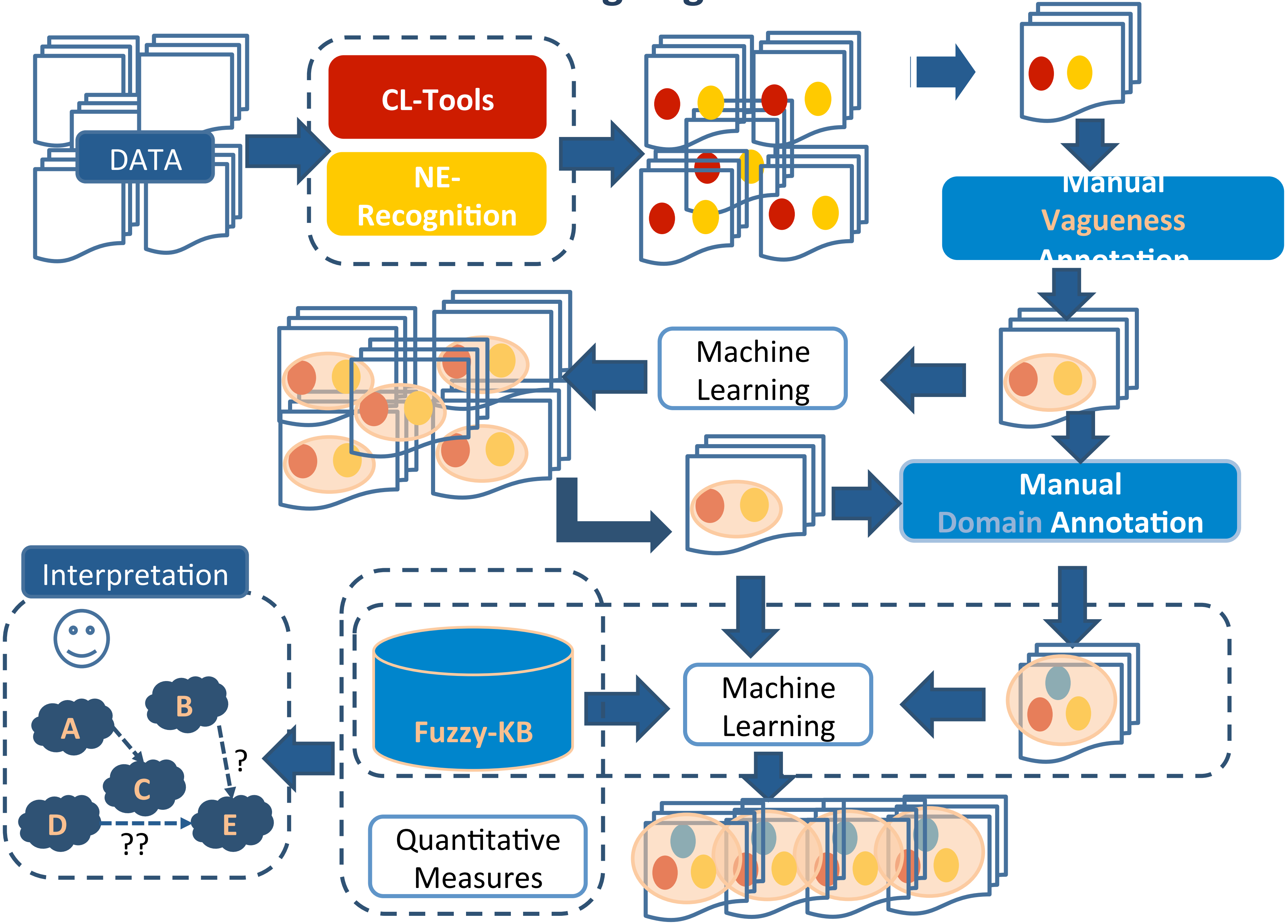
decreasing credibility

historical distance
- modern
- historical

Current DH - Approach



Including Vagueness



DATA

CL-Tools

NE-Recognition

Manual Vagueness Annotation

Machine Learning

Manual Domain Annotation

Interpretation

Fuzzy-KB

Machine Learning

Quantitative Measures

Lexical and Syntactic Sources of vagueness in the original

Quotation

Hactenus Gregoras: ad cuius verba observare haud extra propositum erit τὴν πρώτην, quam Gregoras vocat „**Tart**[illegible]dem esse, quam hodie vulgo [illegible]mus eiusque incolarum no[illegible] historicis recenseantur, tamen adscita magis, aut ab exteris indita, quam propria eisque, dum in suis sedibus [illegible]uliaria fuisse. Ita, si qui[illegible]oni Praefatione legerit O[illegible] Principes in duas stirpes fuisse divisos, „Aliothman” unam, et „Ali Dzeng[illegible] credat sub ipsis horum ge[illegible]hanc appellationem[illegible] eas gentes invaluisse. Vti enim absonum videtur, Aliothmanos Suleimano parentes ab h[illegible]integro post saeculo iis imperab[illegible] sortitos; ita non minus falso vulgo praedicantur Tartarorum Crimensium Principes ab ipso Dzengizchano „**Alidzengiz**” appellationem retinuisse.

More plausible

Would have been…

seems unlikely

equaly false

Până aici l-am citat pe Gregoras: față de cuvintele lui nu va fi nepotrivit să observăm că acea Tartaria „ἡ πρώτη”, pe care o numeşte Gregoras, este chiar aceea pe care o numim îndeobşte cea „Mare”, iar numele locuitorilor ei, chiar dacă sunt înregistrate de istorici, au fost totuşi mai degrabă împrumutate sau date de străini decât proprii lor, purtate întocmai pe vremea când se aflau în sălaşurile lor. Astfel, dacă va fi citit cineva în Prefața pusă înaintea acestui tratat că principii neamului oguzilor au fost împărțiți în două stirpe, una „aliothmană”, cealaltă, „alidzengiză”, să nu creadă că denumirea aceasta era de-acum valabilă pentru întemeietorii acestor neamuri. Căci, după cum pare nepotrivit ca aliothmanizii care i se supun lui Suleiman să-şi fi ales numele de la nepotul acestuia, care a domnit peste ei după un secol întreg, la fel de fals se spune îndeobşte că principii tartarilor din Crimea şi-ar fi păstrat denumirea „alidzengiz” chiar de la Dzengizchan

Example for wrong knowledge extracted without deeper linguistic annotation – German and Romanian case

Domnul cel dintâi carele după năvălirea lui Batie, a agonisit iarăși strălucirea cea mai dinainte a Moldovei a fost:
1. Dragoș și măcar că hronog oastre nu arată pentru știința neamul i sau, dar la noi se zice necontenit, că **a fost** din neamul cel vechiu al crailor Moldovinești, și a avut tată pe Bogdan fiul lui Ioan, dela carele toți Dom-nii obișnuesc a-și pune la iscălitură numele Ioan.

was

Si cuvântul acesta este mai usor de a se te

Dragos = belongs_to Moldavian kings

de a se crede, că altul din neam mai prost, ar fi putut cu o tovărășie așa mare să meargă la vânat, carele a dat prilej la descoperirea Mol-dovei și ar fi putut …

Der erste demnach, der nach Batia Einfall (*) der Moldau ihren vorigen Glanz wieder verschafft hat, war
1. Dragosch. Obgleich unsre Jahrbücher sein Geschlechtsregister nicht a so ist es doch eine beständige S uns, daß er aus dem alten königlich moldauischen Stamme **gewesen sey**, und den Bogdan zum Vater gehabt habe,

should have been

Dragosch ≈ belongs_to Moldavian kings

welchem alle Fürsten den Namen Johannis in ihrem Titel zu führen pflegen; dieser Meinung ist desto mehr Glauben beyzumessen, weil man schwerlich glauben kan, daß einer von gemeiner Herkunft mit einem so großen Gefolge auf die Jagd (welche die Moldau zu entdecken Gelegenheit gegeben,) habe ausgehen, ….

Example for necessary Manual Annotation of Factual uncertainity

[…] He fought two Battles with Bajazet Ildirim; in the first he was victor, and in the second he routed him with a memorable slaughter, which seven vast piles of *Turkish* Bodies erected after the Battle, witnessed, by the Confession of *Hezarfenn* himself, the faithful *Turkish* Historian.

Cantemir, pp. 47 (Annotations)

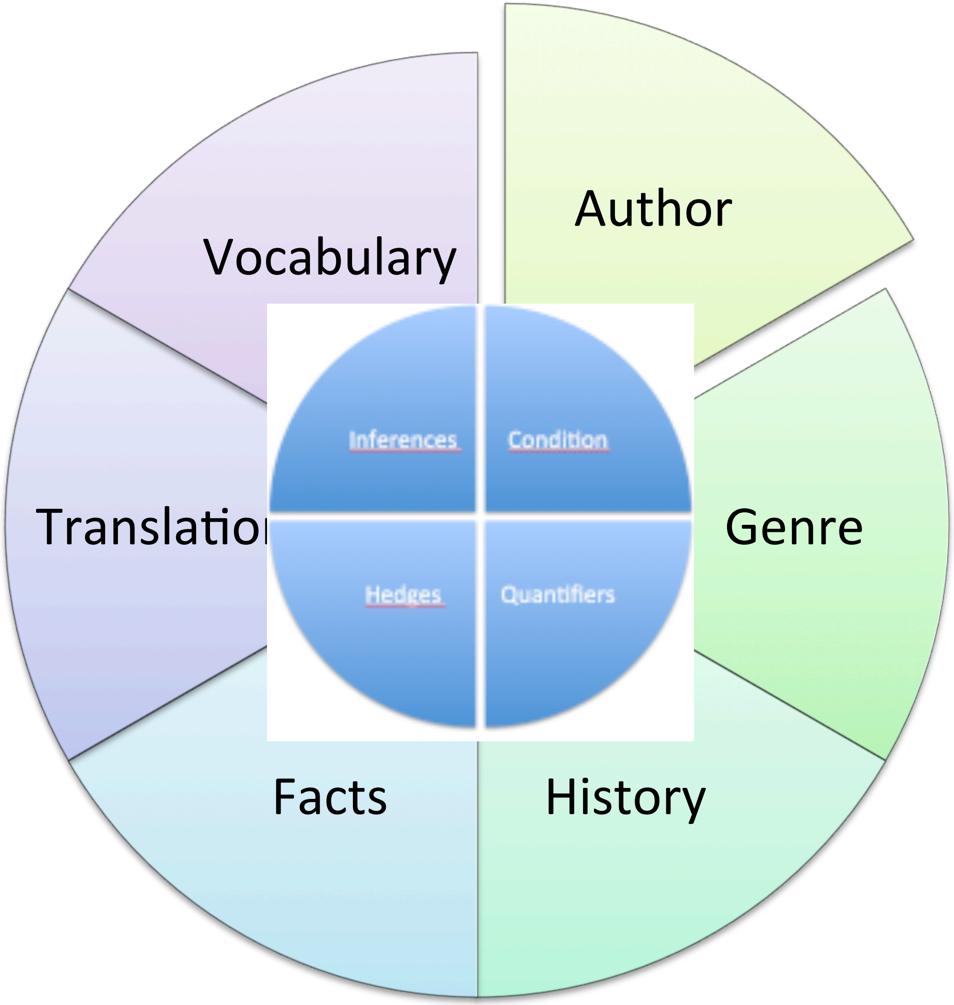
Hezarfen (Hezarfen Hüseyin Efendi) (?-1691/92), Tenkih-i Tevarih-i Mülük: is **NOT** mentioning these facts

The Turkish historians so extoll this prince's expedition in assembling his troops, in executing his designs, and in vanquishing his enemies, that when they talk of the natural speed of the Tartars in comparison with his wonderful marches, they call the first, the creeping of a snail.

Cantemir, pp. 48 (Annotations)

Described in Solakzade: ?, Hoca Saadettin: , Neşri:

How to represent vagueness



Summary

To avoid,

- that words/texts become facts or concepts without semantic annotations,
- that big social data become uniform data base entries without some sort of reliability check,

we need indications of their vagueness.

References

- Thomas T.Ballmer and Pinkal, Manfred, Approaching Vagueness, Amsterdam 1983
- Geeraerts Dirk, Vagueness's puzzles, polysemy's vagaries. In: Newman, John Cognitive lLinguistics. Berlin 2009.
- v.Hahn, Walther, Vagheit bei der Verwendung von Fachsprachen. In: Hoffmann / Kalverkämper /Wiegand: Fachsprachen. Band 1. Berlin 1998. S. 383 – 390.
- Pinkal, Manfred, Semantische Vagheit: Phänomene und Theorien, Teil I. In: Linguistische Berichte Nr. 70, S. 1-26, Wiesbaden 1980.
- Pinkal, Manfred, Semantische Vagheit: Phänomene und Theorien, Teil II. In: Linguistische Berichte Nr. 72, S. 1-26, Wiesbaden 1981.
- Edeltraud Winkler, Überlegungen zu Artefaktbezeichnungen im Deutschen. In: Deutsche Sprache 37 (2009) H. 1, S. 33-47.