Automatic Stemma Generation

Circumnavigating the Data Sparseness Problem by Means of Random Text Generation and Copy Process Simulation

Armin Hoenen

Text Technology Group, Goethe-Universität Frankfurt am Main

July 4, 2011

Corpora

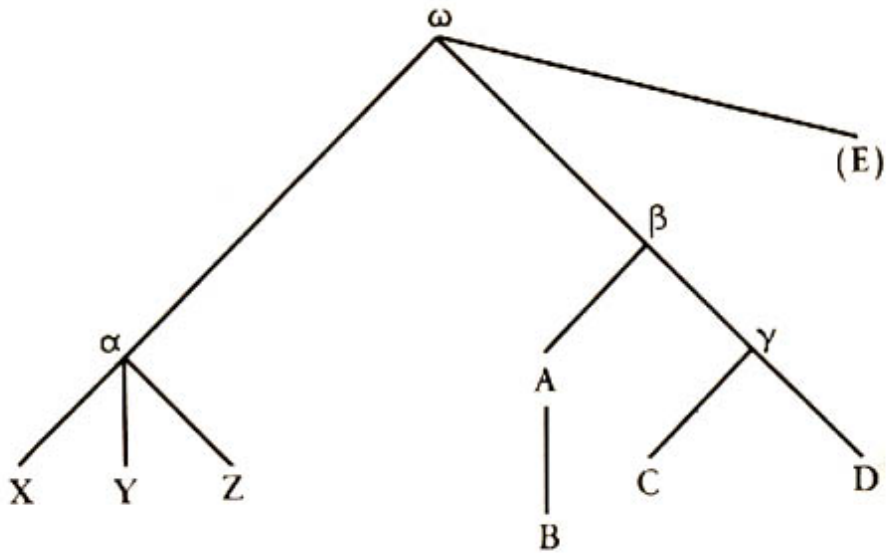
TITUS Archives[1]:

- Avestan Corpus (4; 7 MS)
- Old- Georgian Gospels (40 MS)

[1] http://titus.uni-frankfurt.de

Armin Hoenen

Stemma Codicum



[2]

Stemma codicum: Directed affiliation graph of ancient manuscripts.

[2] *http://1.bp.blogspot.com/_cA728IVqkJU/S7WRr2PSkmI/AAAAAAAAAxQ/xS-T-Zw9deM/s1600/stemma+codicum.jpg*

Copy Process

a copy process introduces **variants**

3

3 http://blog.drewberman.com/perfect-business/mlm-training/mlm-mistakes/

Goals

1. Which manuscripts are the oldest?
2. Which text version is most authentic?

Previous Work

ASG is available for several medieval manuscript corpora:

1. Canterbury Tales (Robinson et al. 1998)
2. Lanseloet van Denemerken (Salemans 2000)
3. St. Henry of Finland (Roos et al. 2005)
4. Slavonic gospels (Mironova 1996)

Armin Hoenen

Analogy to Biology [1]

*Manuscript Text (**MS**) <=> **DNA***

Benefit: Programs for automated stemma generation exist, but we find *text and copy process specific characteristics*

Similarities

Similarities between Genes and Manuscripts:

Manuscripts	**DNA**	**Agreement**
typographic mistake	punctual mutation	Y
abbreviations, omissions, deletion	deletion	(Y)
added comments/sections ...	Gene insertion	Y
contamination	lateral gene transfer	Y
error correction	backmutation	Y
scrambled passages	crossing- over	(Y)
alphabetic	4 letter code(CGAT)	Y

- Programs from bio-informatics can be used ([1][2])
- more information in less data
- a lot more possible sequences of length n from longer alphabet(!)

Similarities

Similarities between Genes and Manuscripts:

Manuscripts	DNA	Agreement
typographic mistake	punctual mutation	Y
abbreviations, omissions, deletion	deletion	(Y)
added comments/sections ...	Gene insertion	Y
contamination	lateral gene transfer	Y
error correction	backmutation	Y
scrambled passages	crossing- over	(Y)
alphabetic	4 letter code(CGAT)	Y

- Programs from bio-informatics can be used ([1][2])
- more information in less data
- a lot more possible sequences of length n from longer alphabet(!)

Similarities

Similarities between Genes and Manuscripts:

Manuscripts	DNA	Agreement
typographic mistake	punctual mutation	Y
abbreviations, omissions, deletion	deletion	(Y)
added comments/sections ...	Gene insertion	Y
contamination	lateral gene transfer	Y
error correction	backmutation	Y
scrambled passages	crossing- over	(Y)
alphabetic	4 letter code(CGAT)	Y

- Programs from bio-informatics can be used ([1][2])
- more information in less data
- a lot more possible sequences of length n from longer alphabet(!)

Algorithms

In biology the state of the art methods([1]) are:

Methods

1. Distance Matrix
2. Parsimony
3. Maximum likelihood
4. Split decomposition
5. others

In stemmatology **parsimony** and **split decomposition** are used most.

Tree Typology I

The different algorithms produce different types of trees.

The distinction between

- *rooted* and *unrooted*
  and
- *bifurcating* and *multifurcating*

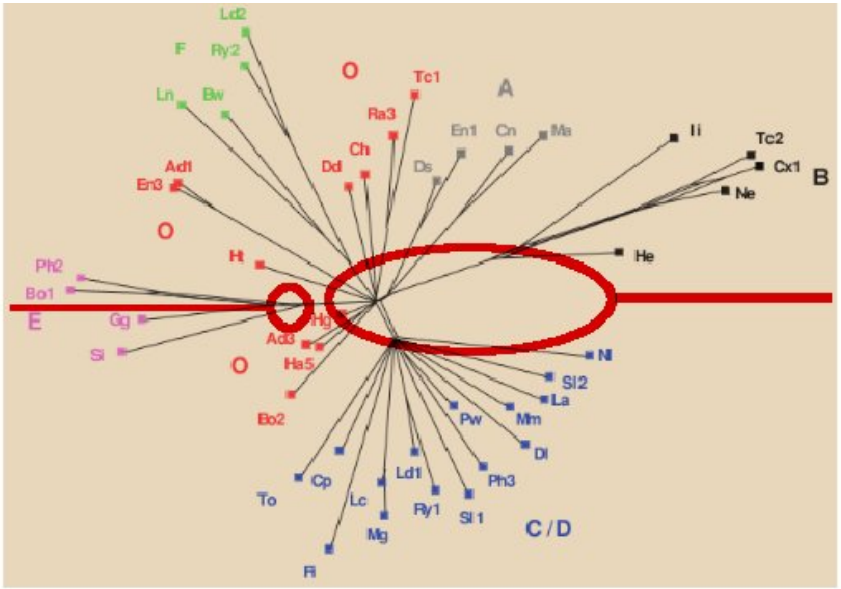
is what is most important in our case.

Tree Typology II

We need unrooted multifurcating trees, because we don't know which one is the oldest manuscript (**root = ?**)

multiple copies from one manuscripts are highly probable (**» multifurcating**)

Our Desired Tree



4

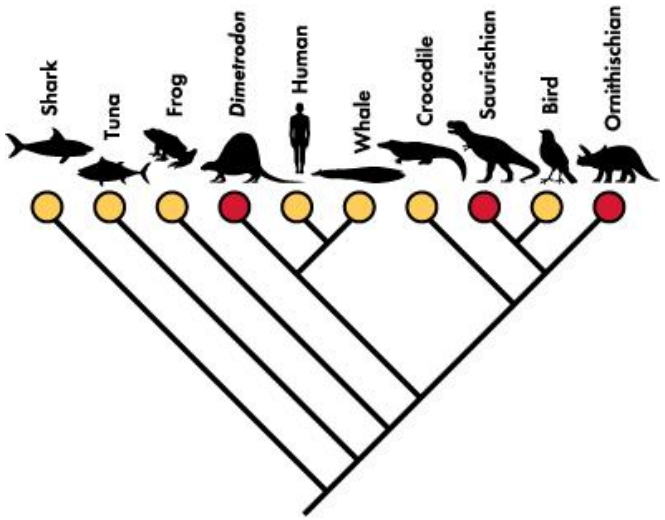
Figure: multifurcating unrooted

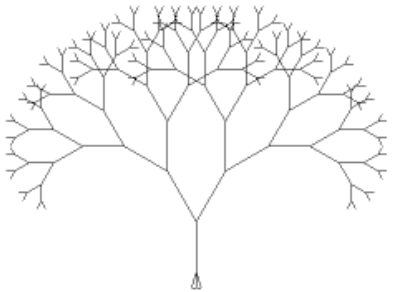
[4] Img from: Robinson et al. In NATURE 394 27.8.1998

Other Trees (Firewood)

Armin Hoenen

rooted bifurcating trees





Number of Trees

Stemma computation is intense

Possible Stemmata [4]

1. $2n!/n!(n+1)!$ for binary trees
2. Avestan (small):4 » 14
3. Avestan (big):7 » 429
4. Georgian: 15 + 1/ month: 30 » $3,814,986,502,092,304$

Best tree = **NP hard problem**

Metadata Restriction

For computation one should restrict number of trees if possible.

Colophons are prolegomena or postscriptae, short notes, containing metadata, such as:

- manuscript copied from (!!!)
- year, place and orderer of copy
- name of copyist
- ...

Colophons yield a **known stemma fragment**: use this as restriction for tree computation.

Is stemma generation the only task?

NO!

Is stemma generation the only task?

NO!

Digital Humanities

Digital	**Humanities**
texttechnological analyses	annotation
random text generation (RTG)	variant classification
copy process simulation (CPS)	metadata
automatic stemma generation (ASG)	evaluation of results

Data Sparseness Problem (DSP)

Problems
27925 Avestan tokens (Yasna)
6455 wordforms
3177 hapax legomena

x 4 manuscripts

==========

DSP (Data Sparseness Problem)

DSP++

1/2 of the text consists of repetitions
» religious genre/mnemonic aid

+ helps identify which errors are constant throughout one MS, and which ones are typographic in nature (copyist identification possible ?)

−this reduces the variation among the data considerably» DSP

Approach

There is too few data to train a machine learning stemma generation algorithm.

Approach:

- Create authentic random text of different lengths
- Create corpora of different sizes by simulating a copy process [5]
- ASG

[5] we record each copy relation and get known stemmata for the training

Random Text Generation

Random Text from known Avestan Lexicon should mirror all relevant properties of the original text.

- statistical distributions (frequency, word length, POS sequences ...)
- repetitiveness
- variant distribution

Advantage of random text

Control of these parameters.

Copy Process Simulation

- text level
  - produce variants (weights (0..1))
    - typos
    - dialectal variants/synonyms
    - correction, historical adjustment
  - change distribution of linebreaks (inducing errors)
  - editorial changes (contamination, scrambling, deletion, insertion)
- manuscript level
  - probability l of manuscript loss (from other corpora)
  - prestige value w of manuscripts (prob. to be a mastercopy)

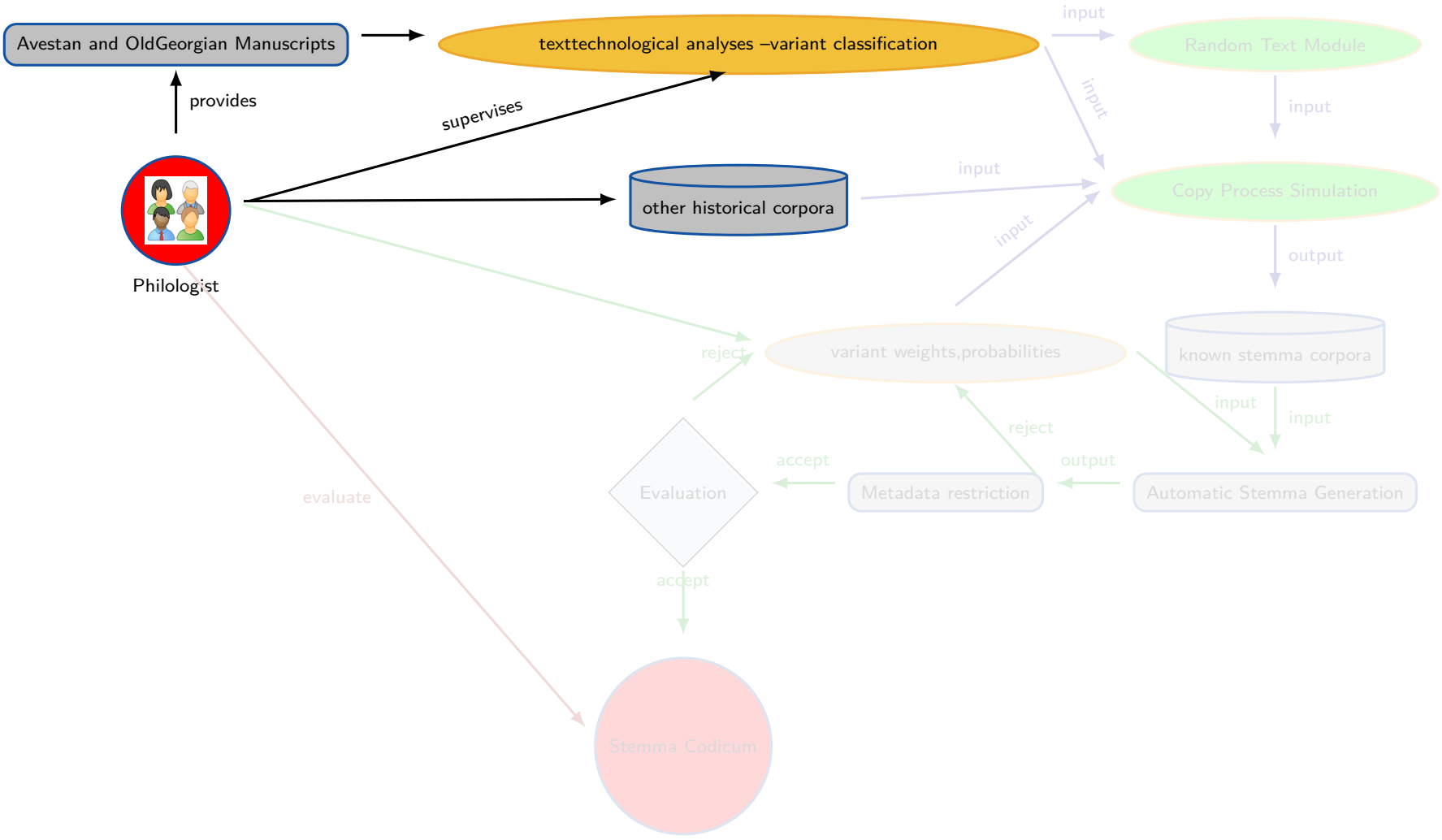
Parameters

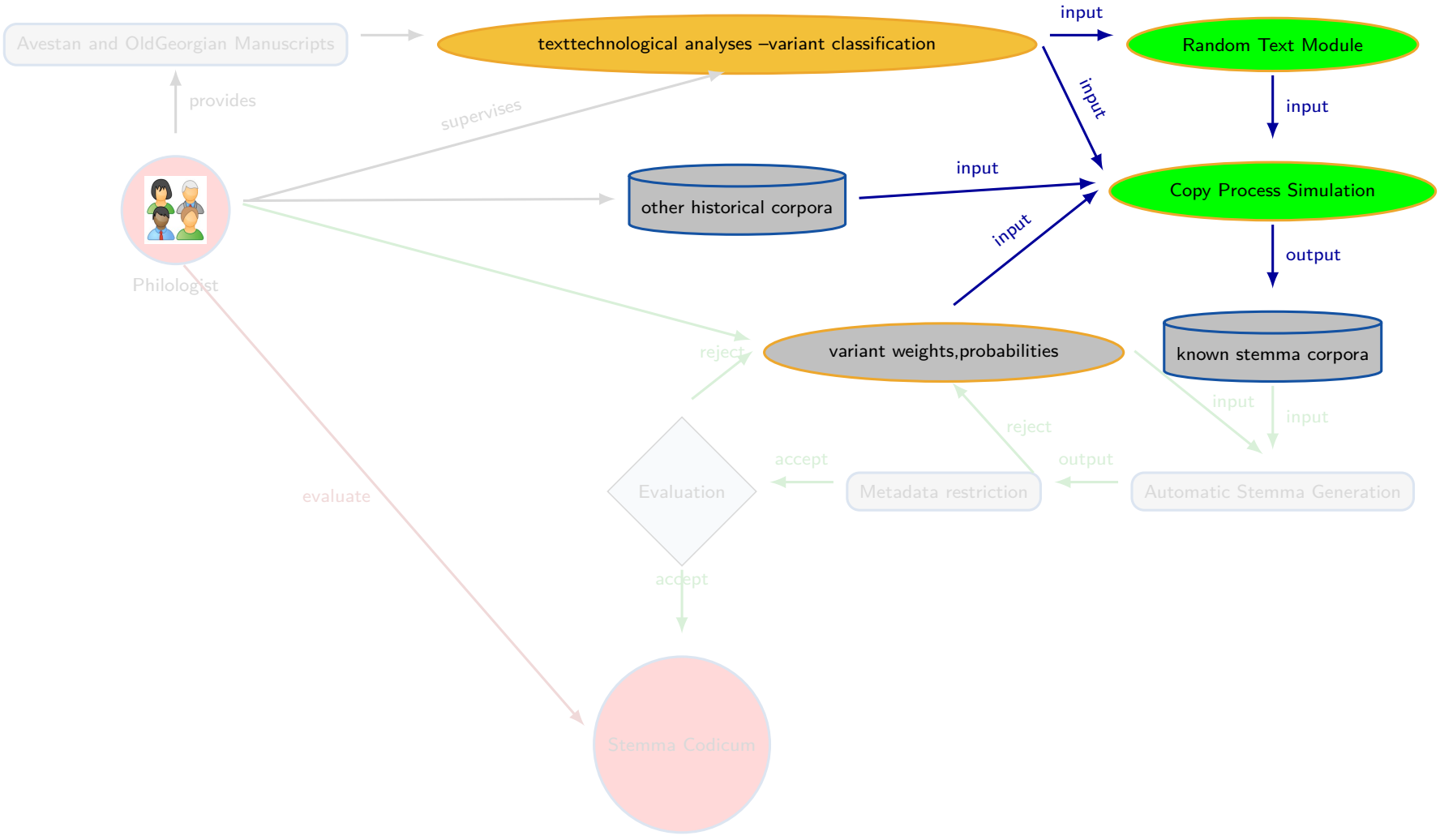
Learning will relate to:

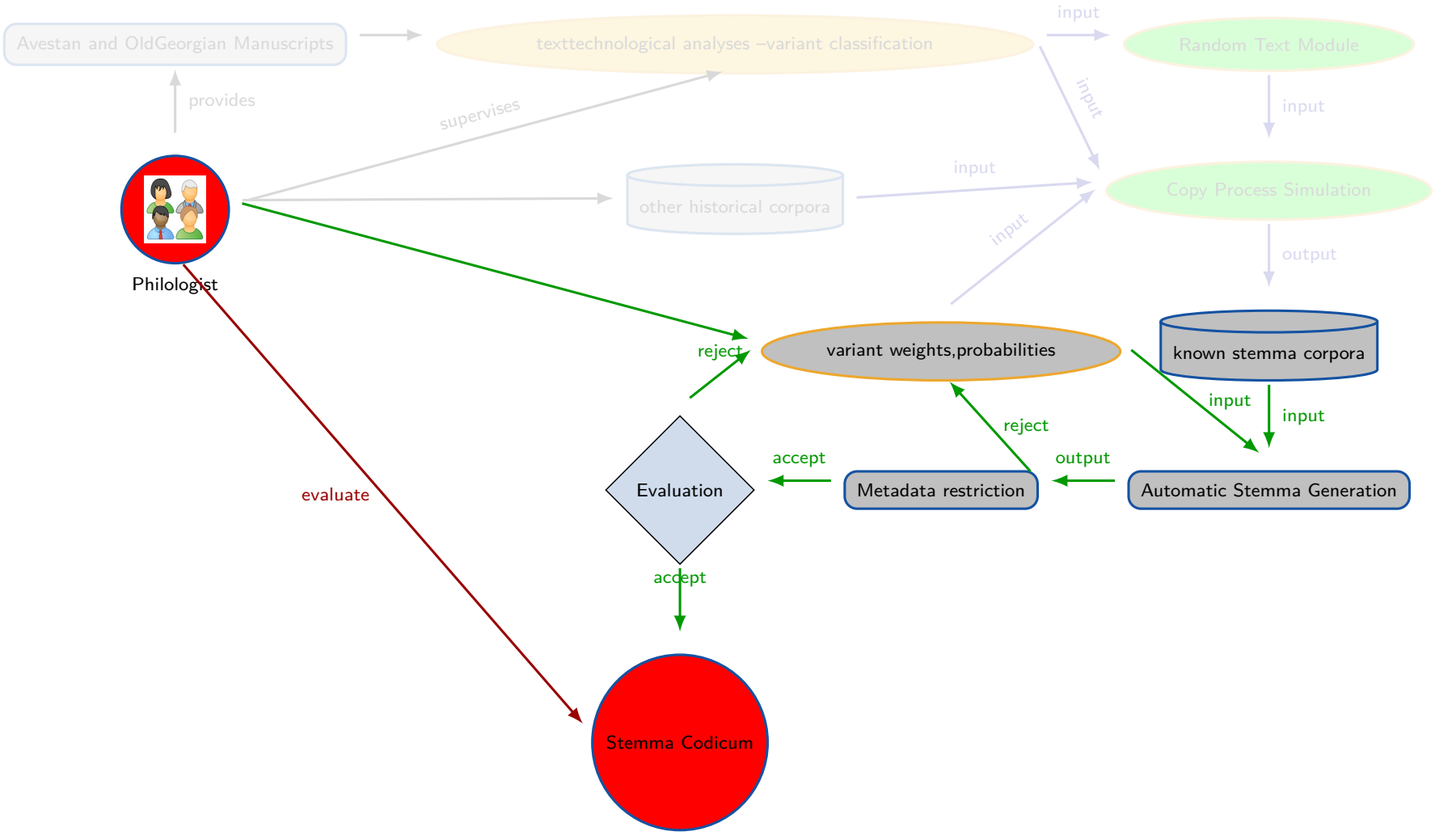
- manuscript length
- corpus size
- weights of variants
- probabilities of MS- loss, MS- prestige

Advantage of CPS:

- Circumnavigation of the DSP (infinite corpora)
- Recording reliability measures


Avestan and OldGeorgian Manuscripts
texttechnological analyses –variant classification
provides
supervises
Philologist
other historical corpora
input
Random Text Module
input
input
input
Copy Process Simulation
output
variant weights,probabilities
known stemma corpora
reject
input
input
reject
accept
output
Evaluation
Metadata restriction
Automatic Stemma Generation
evaluate
accept
Stemma Codicum


Avestan and OldGeorgian Manuscripts
texttechnological analyses –variant classification
input
Random Text Module
input
provides
supervises
input
input
other historical corpora
Copy Process Simulation
input
output
Philologist
reject
variant weights,probabilities
known stemma corpora
input
input
reject
accept
output
Evaluation
Metadata restriction
Automatic Stemma Generation
evaluate
accept
Stemma Codicum


Avestan and OldGeorgian Manuscripts
texttechnological analyses –variant classification
Random Text Module
provides
supervises
input
input
input
input
Philologist
other historical corpora
Copy Process Simulation
input
output
reject
variant weights,probabilities
known stemma corpora
input
input
reject
accept
output
Evaluation
Metadata restriction
Automatic Stemma Generation
evaluate
accept
Stemma Codicum

Step I - Gathering

1. extract metadata
2. classify variants
3. analyse texts statistically

Step II, Step III

- Step II - creating resources
  - RTG ([5][6])
  - **CPS**
- Step III - producing the stemma
  - ASG training
  - ASG on real Data

Summary

Questions assessed:

explicit

1. Stemma for Avestan, Old Georgian
   Which corpora sizes and variant numbers produce how reliable stemmata?
2. Which variant classes are most reliable for stemma generation?

implicit

1. Can we reliably identify a Copyist from individual variant distributions?
2. Can we reliably distinguish 2 translations of the same source text from 2 manuscripts, where one is the copy of the other?

Resources

[1] van Reenen et al. (1996) Studies in Stemmatology. John Benjamins Publishing: Amsterdam.

[2] van Reenen et al. (2004) Studies in Stemmatology II. John Benjamins Publishing: Amsterdam.

[3] http://bioinformatics.istge.it/bcd/Curric/MathAn/node11.html

[4] http://www.durangobill.com/BinTrees.html:20.06.2011

[5] Biemann. 2004. A Random Text Model for the Generation of Statistical Language Invariants.

[6] Ferrer i Cancho et al. 2001. The small- world of human language.

THANK YOU!!