Late Latin Charter Treebank

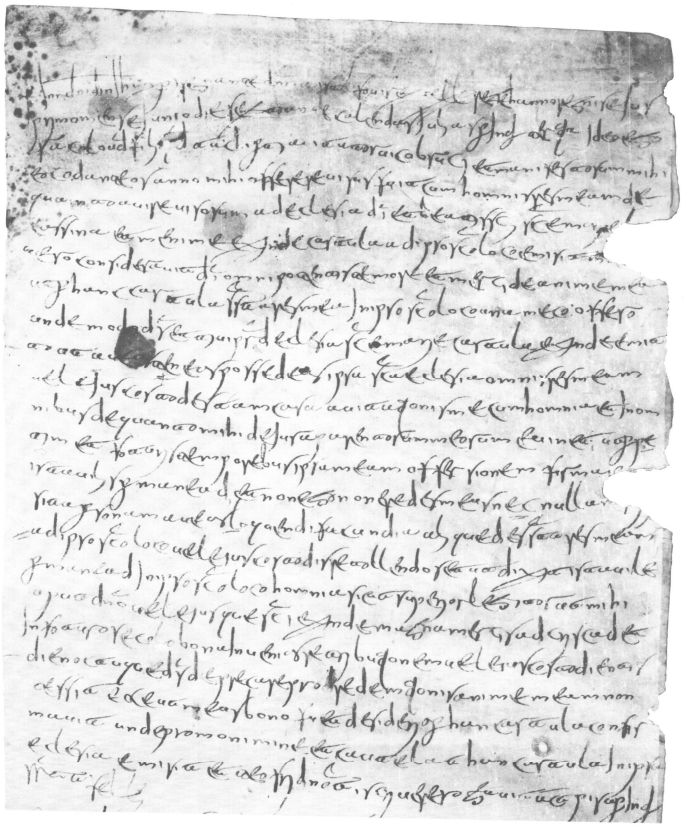
building an extensive dependency treebank of early medieval documentary Latin

Timo Korkiakangas, PhD
Academy of Finland/University of Helsinki
LiLa Workshop *Linguistic Resources & NLP Tools for Latin*
4 June 2019, Milan

Outline

1) charters and their Latin
2) advantages and disadvantages for corpus-linguistics
3) structure of LLCT
4) constructing LLCT1, LLCT2, LLCT3
5) problems and idisoyncracies of LLCT annotation
6) current stage and future perspectives

Charters = private and semi-public documents



Example of charter Latin:

...Ideo constans me Pertus huna cum filius meus Tusculu uiri deuoti binditoris bindedisse et bindedimus bono animus, tadedisse et tradedimus bobis Barbarus et Basus filios et Sad fratis eius terras iuris meis in fundo Cellules, adminsuratas cum aliquantulas bineas factas, quod est totas in circus perticas XXX, et ipsa perticas abente in se per una pedi XX ad pede Munichisi het ipsa terras posita est prope casa bestras emtoris, et de aliam parte, bersura Lupulos. et ipsa terras bobis qui supra emtoris coram testibus tradedimus cum homnis pumi suis, quidquid superpositum habes. et suscipemus nus qui supra binditoris ad bu suprascripti emtoris pro ipsa suprascripta terras cum homnias que supra positas abes integro pretius auri soledus hobridiacus pensantis numerus duo et II trimissi adfenitus; quatenus hab hodierna die abeatis, teneatis, possedeatis, bel, si binderes aut dunares bolueritis, liberas in omnebus habeatis potestatem...

CDL 66, AD 738, Massa Mustiba, Warnegausu notarius

Why (not) charters?

- pros
  - rich non-literary material
  - original documents
  - ‘non-standard’ Latin
  - abundant metadata: scribe name, location, date
- cons
  - formulaicity
  - some linguistic phenomena are continuously repeated, others are absent
  - complicated relation to spoken language: mixture of archaizing, misunderstood, and spoken-language features

Late Latin Charter Treebank

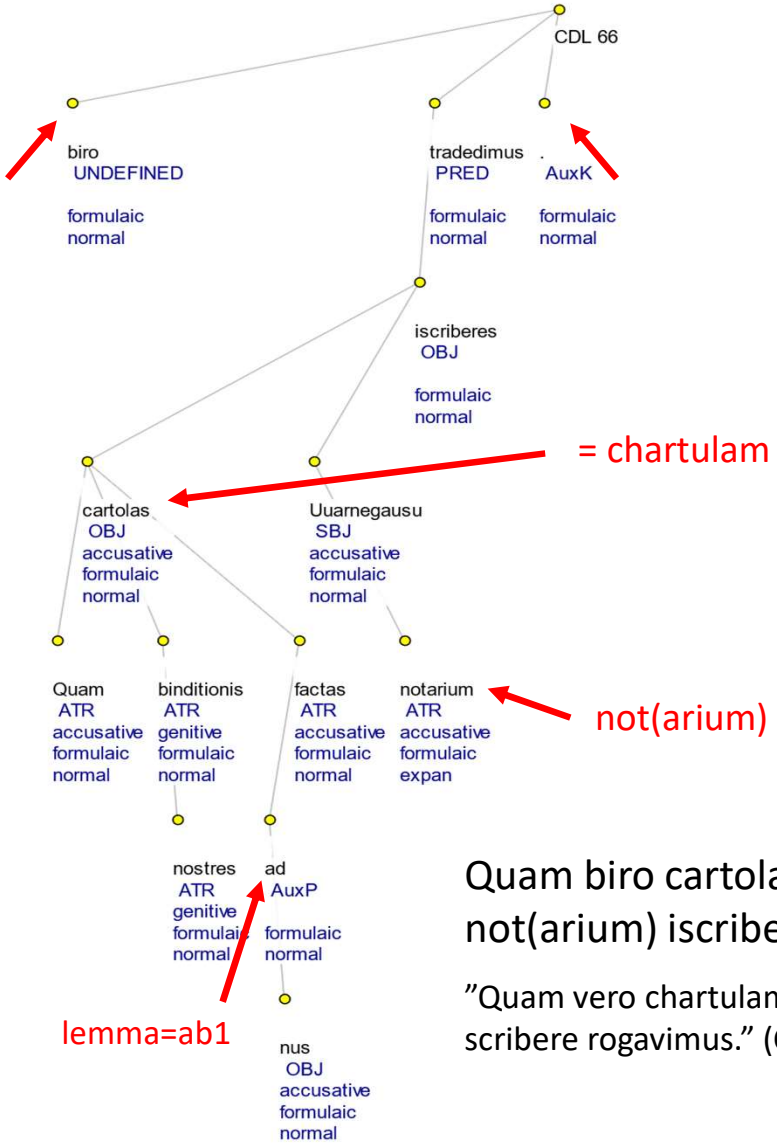
- Italian documentary texts from between AD 714 and 999
- three parts, constructed in three phases beginning from 2010

Treebank part	Constructed	Documents	Tokens	Words	Location	Years
LLCT1	2010-2014	519	225,834	198,696	Tuscany	714-869
LLCT2	2016-2018	521	257,418	220,797	Tuscany	774-897
LLCT3	2019-	~ 217	~ 100,000	~ 85,000	Tuscany, North and South Italy	721-999

Annotation issues

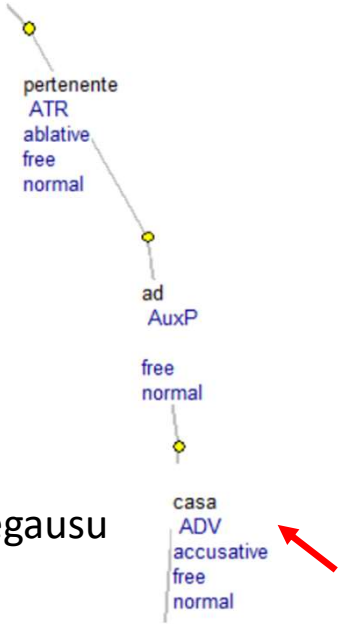
- syntactic annotation based on the *Guidelines for the Syntactic Annotation of Latin Treebanks* (Bamman et al. 2007) – dependency grammar
- lemmatization (and morphological annotation of standard forms) imitates that of the Latin and Ancient Greek Dependency Treebanks, as they were ten years ago (Perseus Digital Library Project)
- morphological annotation of non-standard forms: Korkiakangas & Passarotti 2011
- textual annotation: custom tagset

Annotation issues



pertenente ad casa

”belonging to (the) house”



Quam biro cartolas binditionis nostres ad nus factas Uuarnegausu not(arium) iscriberes tradedimus .

”Quam vero chartulam venditionis nostrae ab nos factam Warnegausum notarium scribere rogavimus.” (CDL 66, AD 738)

LLCT1

Treebank part	Constructed	Documents	Tokens	Words	Location	Years
LLCT1	2010-2014	519	225,834	198,696	Tuscany	714-869

- mainly manual annotation in the 'Perseus annotation environment'
- quality-checked, but still consistency problems, errors…
- non-fragmentary originals or coeval copies from Tuscia, AD 714-869
- based on three copyright-free editions, readings checked against *Chartae latinae Antiquiores* (ChLA)
  - CDL = *Codice diplomatico longobardo* 1–2, ed. L. Schiaparelli (1929–1933), Tuscan charters 714–774 (diplomatic, high quality)
  - CDT = *Codice diplomatico toscano* 2:1, ed. F. Brunetti (1833), non-Luccan charters 775–813 (outdated)
  - MED = *Memorie e documenti per servire all'istoria del Ducato di Lucca* 5:2, ed. D. Barsocchini (1837), Luccan charters 775–869 (outdated)

LLCT2

Treebank part	Constructed	Documents	Tokens	Words	Location	Years
LLCT2	2016-2018	521	257,418	220,797	Tuscany	774-897

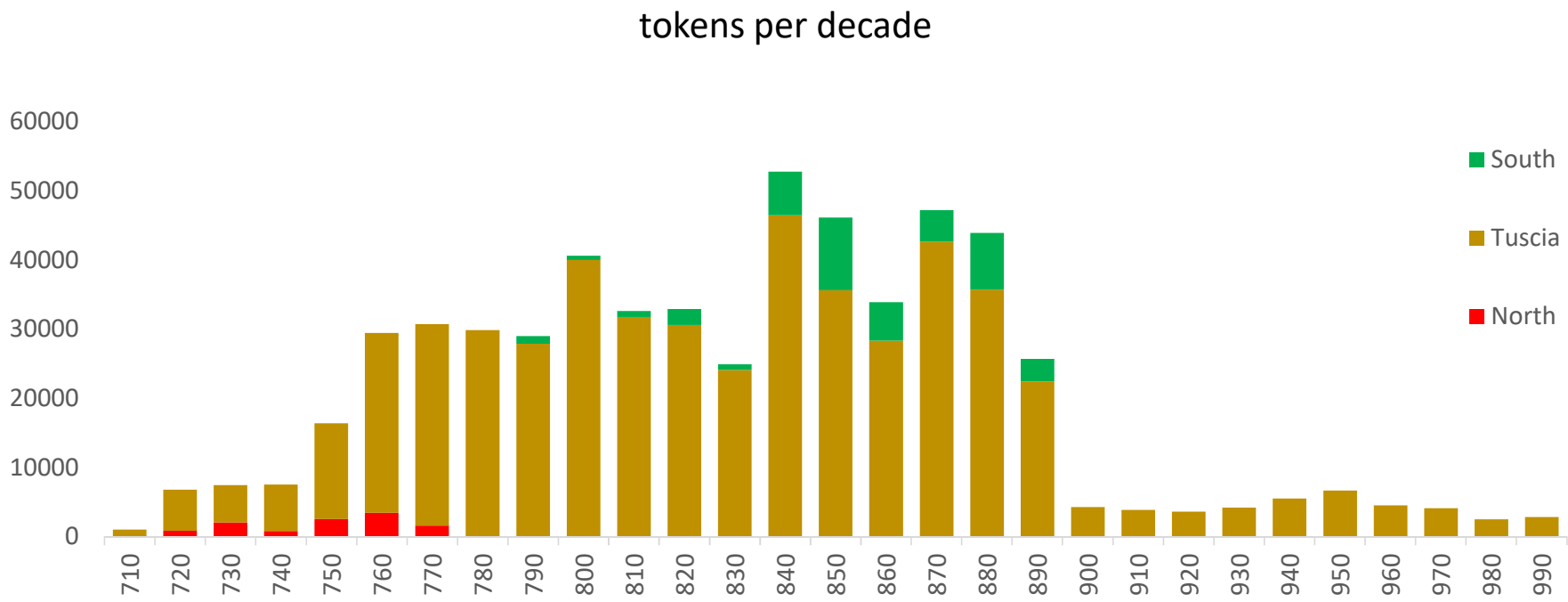
- still Tuscia; chronological extension up to AD 897
- lemmatization: string match on LLCT1
- morph tags: TnT tagger on LLCT1
- syntax: MaltParser on LLCT1
- textual annotation: manual

LLCT3

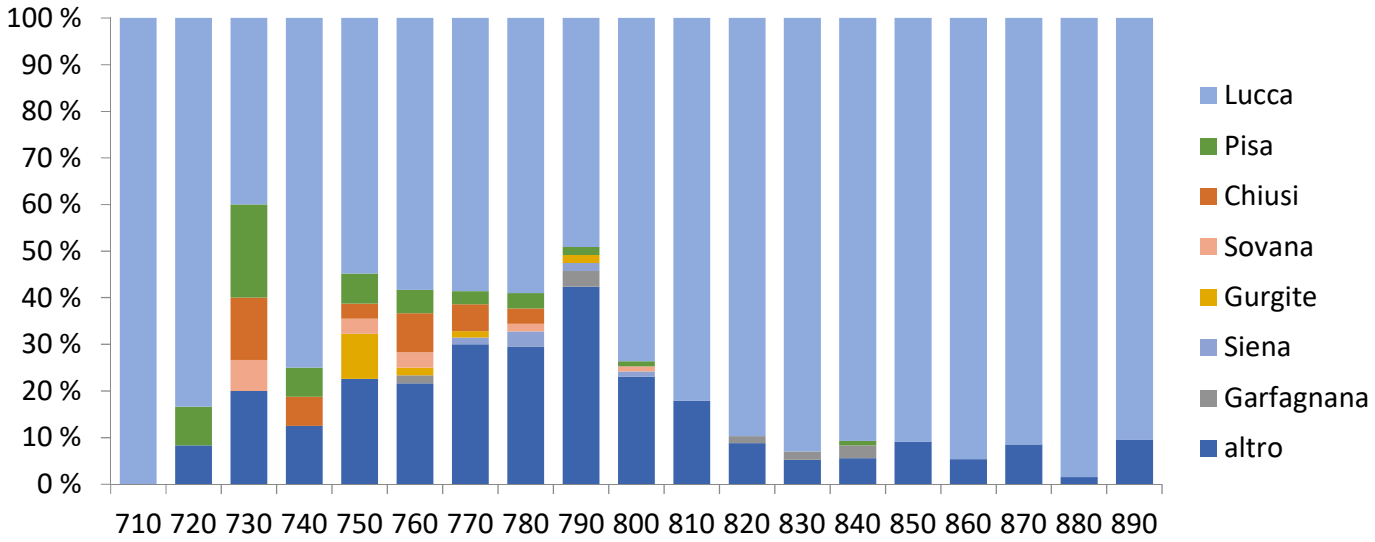
Treebank part	Constructed	Documents	Tokens	Words	Location	Years
LLCT3	2019-	~ 217	~ 100,000	~ 85,000	Tusc., N.+S. Italy	721-999

- chronological and geographical extension
- three datasets under construction:
  1) Lucca, AD 900-999, 73 charters, > 41,871 tokens (transcribed in the archive)
  2) North Italy, AD 721-774, 33 charters, ~11,212 tokens (available in CDL)
  3) South Italy (Cava dei Tirreni), AD 792-899, 111 charters, ~44,136 tokens (available in *Codice diplomatico cavense* 1, ed. M. Schiani & al. 1873 (CDC))
- will be annotated using LLCT2 as the training set + manual textual annotation

Chronology and geography



LLCT1+2: geography



geographical distribution of LLCT1+2
(exclusively Tuscia)

Current stage

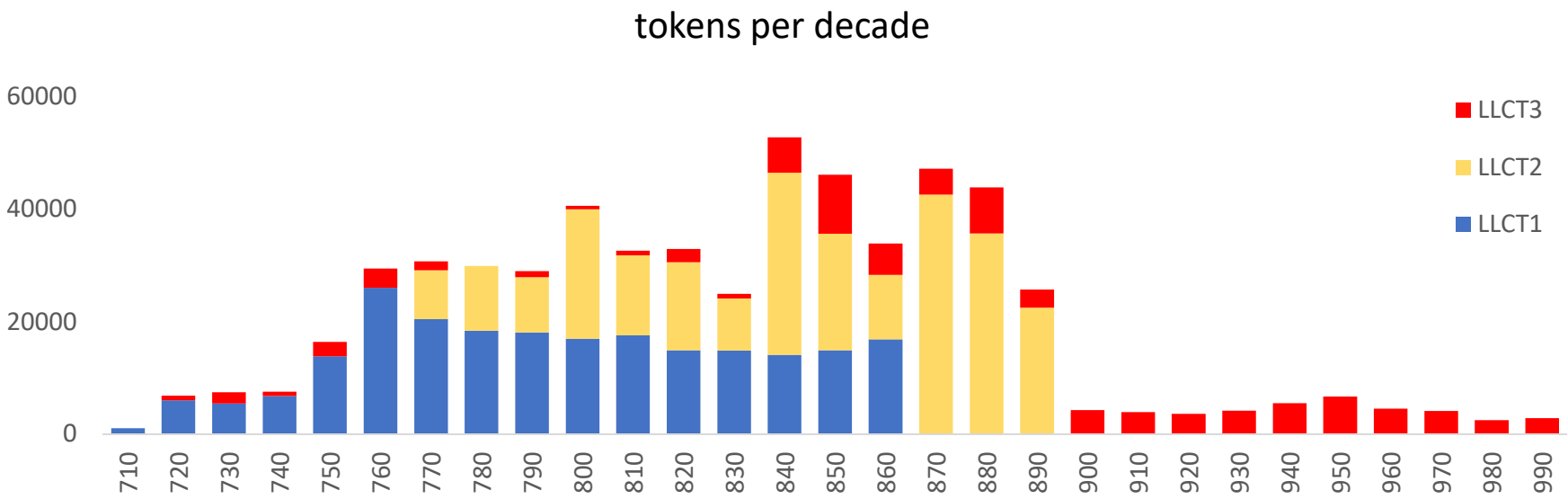
- preparing the corpus for LLCT3
- final manual correction of LLCT2
- preparing UD conversion and merger of LLCT2 to the LiLa Knowledgebase (with help of the LiLa team at CIRCSE)

(for the present, the most up-to-date version of LLCT1 is available at Zenodo in PML format: `doi.org/10.5281/zenodo.1197357`)

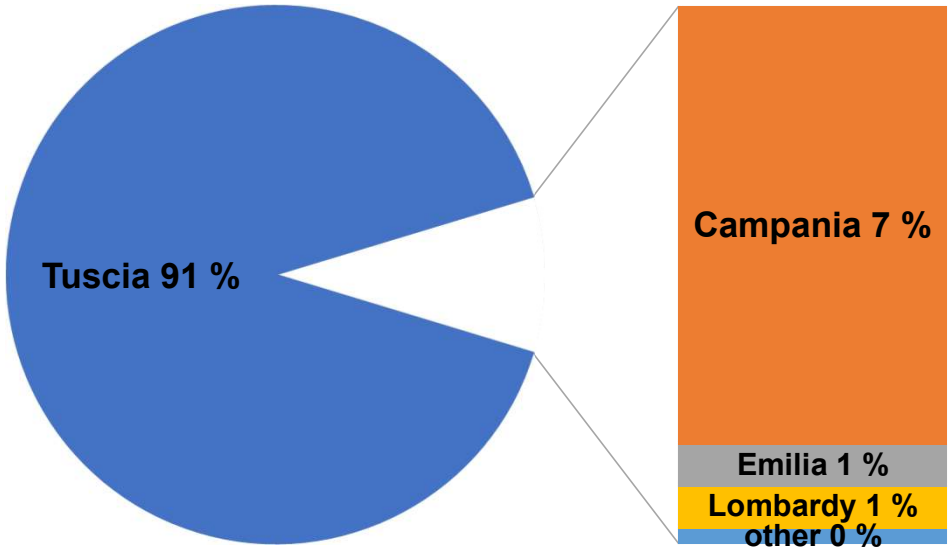
Future perspectives

- completing LLCT3 in 2021 (by the end of my current project)
- merging LLCT1, LLCT2, and LLCT3 into a unified treebank, preferably in UD
  - correcting LLCT1 annotation
- turning other Late Latin charter corpora into treebanks:
  - Merovingian charters (based on PaLaFra-LAT)?
  - Ravenna papyri?
  - Tablettes Albertini?

Chronology

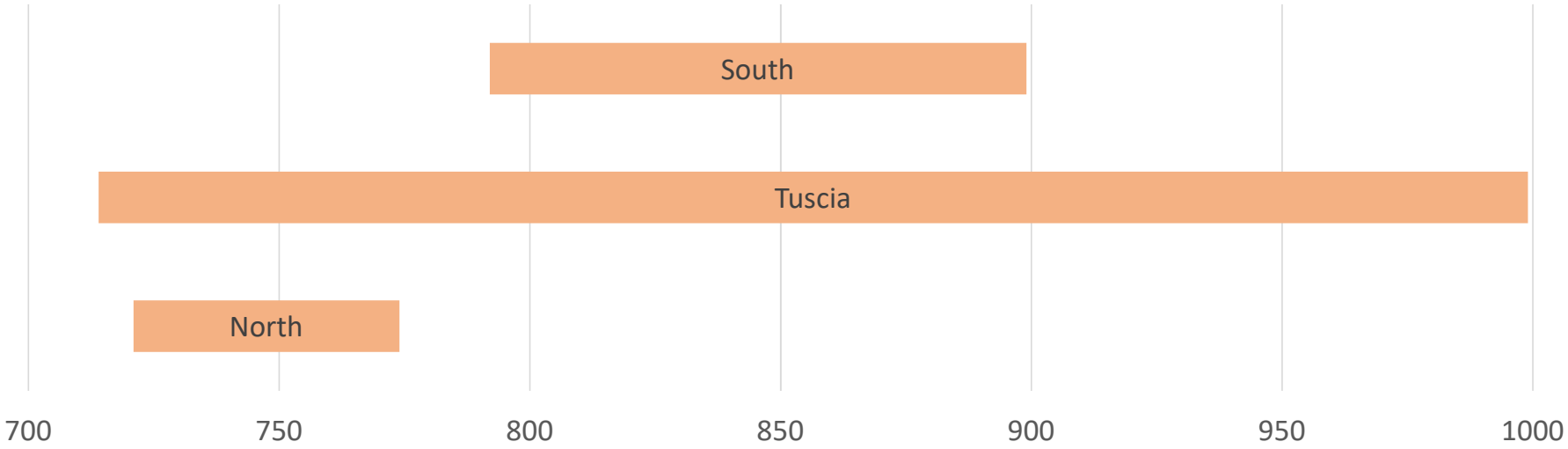

tokens per decade
60000
40000
20000
0
LLCT3
LLCT2
LLCT1
710
720
730
740
750
760
770
780
790
800
810
820
830
840
850
860
870
880
890
900
910
920
930
940
950
960
970
980
990

Geography

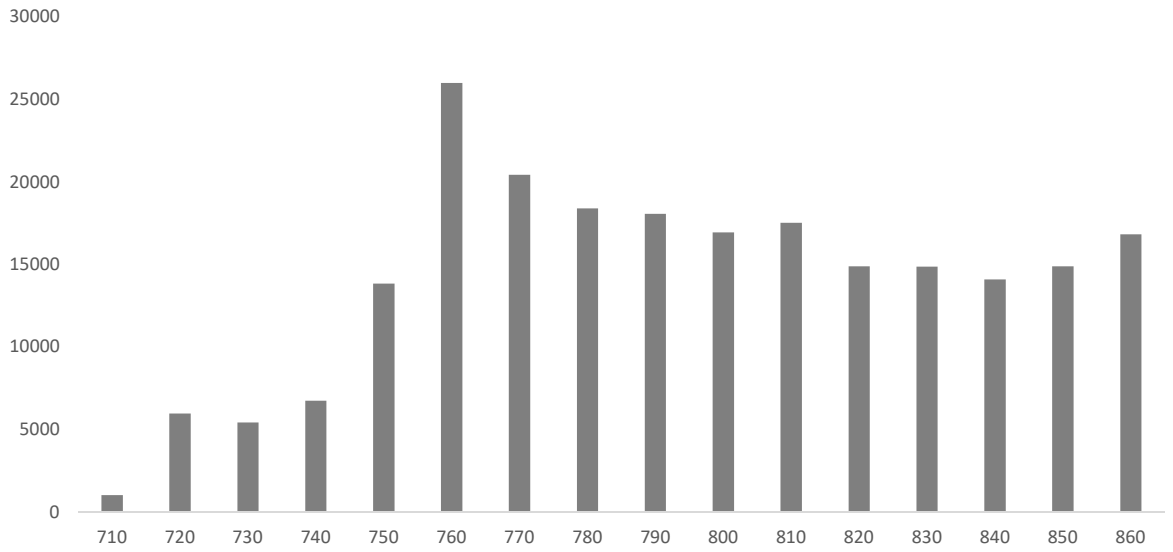


geographical distribution of LLCT1+2+3

Chronology and geography

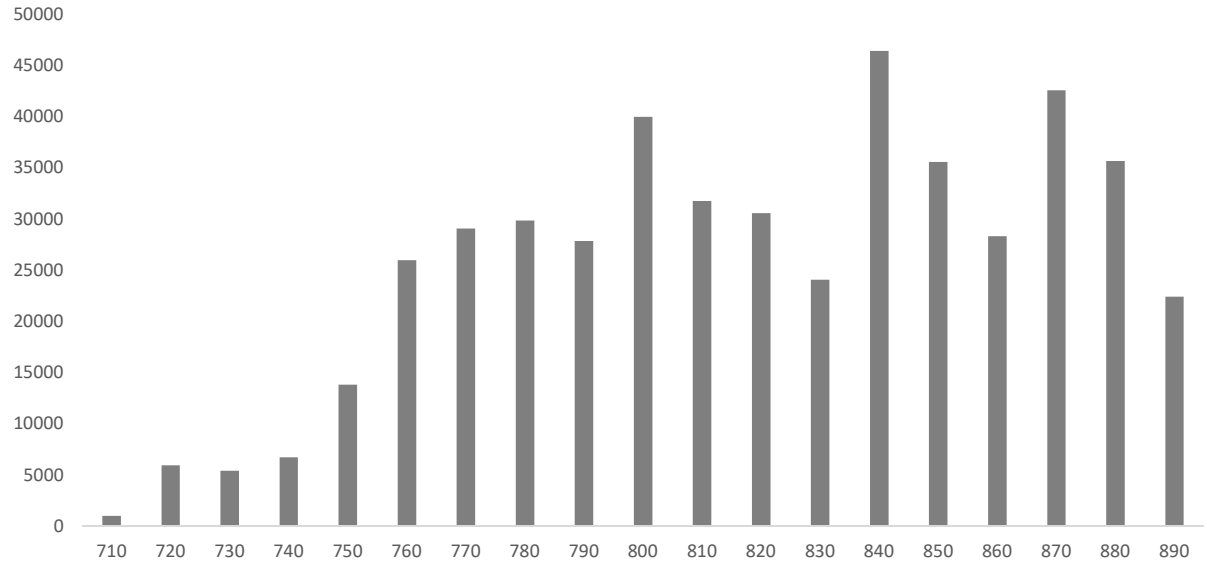


LLCT1: chronology



chronological distribution: tokens per decade

LLCT1+2: chronology



chronological distribution: tokens per decade