Stamtræer og digitale etymologiske ordbøger

Seán D Vrieland

Presentation of etymological information in dictionaries varies widely, making it difficult to propose a standard form of encoding this information in the process of digitization. The following proposes an extension of the TEI P5 standard for encoding texts using XML, wherein a new element <etymon> is able to nest recursively, thereby allowing the XML format to match that of a language family tree. Problems of reconstructed proto-forms, loanwords, and language grouping are further discussed.

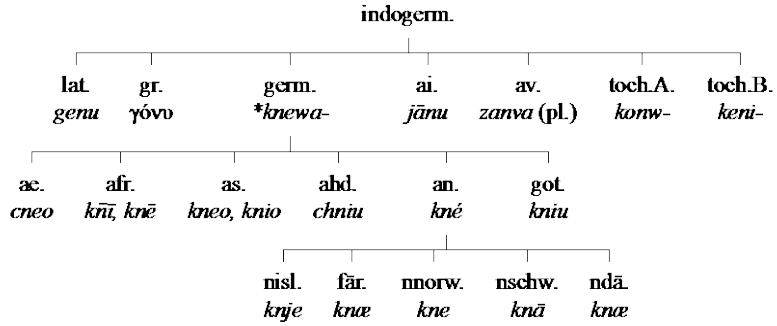
1. Digitale ordbøger og etymologier

Når ordbøger skal digitaliseres, opstår der problemer med at kunne opmærke etymologiske oplysninger på en måde, så de kan læses både af brugeren og computeren. Brugeren vil for eksempel opfatte kognaterne for oldnordisk *kné* i de Vries' (1961) etymologiske ordbog på en helt anden måde, end computeren gør.

> **kné** n. 'knie; glied; krummholz' (< germ. **knewa*-), nisl. *knje*, fär. *knæ*, nnorw. *kne*, nschw. *knä*, ndä. *knæ*. – got. *kniu*, ae. *cneo*, afr. *knī, knē*, as. *kneo, knio*, ahd. *chniu*. – lat. *genu*, gr. γόνυ, ai. *jānu*, av. *zanva* (pl.), toch.A. *kanw*-, B. *keni*- 'knie' (s. über das verhältnis der formen Petersson SVS, Lund, 1, 1921, 7).

For computeren er det muligt at opmærke alle kognaterne som beslægtede med *kné*, men brugeren forstår disse informationer anderledes. Der er stor forskel på den fællesgermanske form **knewa*-, som er en rekonstrueret urform – den oldnordiske forms forfader – og dansk *knæ*, der er en af den oldnordiske forms efterkommere.

Brugeren vil også opfatte tankestregerne i de Vries' ordbog som adskillelsestegn. De nordiske former, der står foran den første tankestreg, og som er opslagsordets efterkommere, er beslægtet med *kné* på en anden måde end de efterfølgende former. Disse former er endvidere placeret i to grupper: germanske kognater, som er tæt beslægtede med *kné*, og ikke-germanske, indoeuropæiske former, som også er beslægtede, men længere tilbage i tiden.



Figur 1: Oldnordisk *kné* i et stamtræ.

En bruger, der har forstået den måde, de Vries (1961) viser de etymologiske informationer på, vil anbringe *kné* og dets kognater i et slags stamtræ som i Figur 1. Spørgsmålet er altså, hvordan disse kognater kan opmærkes med præcis samme struktur.

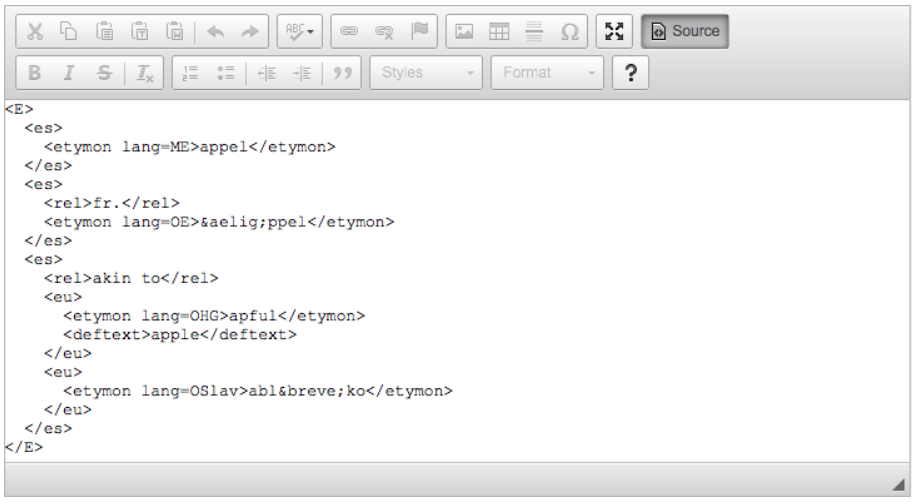
1.1. Tidligere løsninger til opmærkning af etymologier

Sean Crist (2005) diskuterer teoretisk, hvordan etymologiske informationer kunne opmærkes. Han skelner mellem tre opmærkningstyper:

> Type I. Markup schemes which make no provision for etymological data.
>
> Type II. Markup schemes where etymological data is delimited as such, but treated as unstructured prose

> Type III. Markup schemes where the mathematical relationships recognized in historical/comparative linguistics are somehow embodied in the markup system in (semi-)machine-readable form

Som eksempel på type III diskuterer Crist et forslag fra Amsler & Tompa (1988), der betjener sig af et antal SGML-tags til opmærkning af ordbøger. Etymologier i Amsler & Tompas system grupperes i et tag <E> (*etymon*), der indeholder tagget <es> (*etymological segment*) som "branches of a universal etymology tree" (Amsler & Tompa 1988:7) – dvs. ordets slægtskab med en kognat – og <eu> (*etymological unit*) til kognaten og dens betydning. Hver kognat tagges i <etymon> og forsynes med sprogets navn som attribut (for eksempel <etymon lang=OHG> for den oldhøjtyske kognat), og informationer om, hvordan kognaterne er beslægtede, tagges i <rel> (*relation*). Et eksempel fra Amsler & Tompa på etymologien af engelsk *apple* er gengivet i Figur 2.



Figur 2: Etymologien til engelsk *apple* ifølge Amsler & Tompa (1988:4).

Amsler & Tompas forslag giver mulighed for at opmærke kognaterne samt deres definitioner og nogle informationer om, hvordan de er beslægtet med opslagsordet. Men computeren ville ikke kunne læse denne kode og vise kognaterne som et stamtræ, hvor det moderne engelske ord kommer fra det middelengelske, der selv kommer fra det oldengelske. Der gives heller ikke nogen informationer om, hvordan de oldhøjtyske og oldslaviske former er beslægtet med hinanden, herunder ikke mindst hvordan de er beslægtet med opslagsordet.

1.2. Digitalisering af ordbøger ifølge TEI

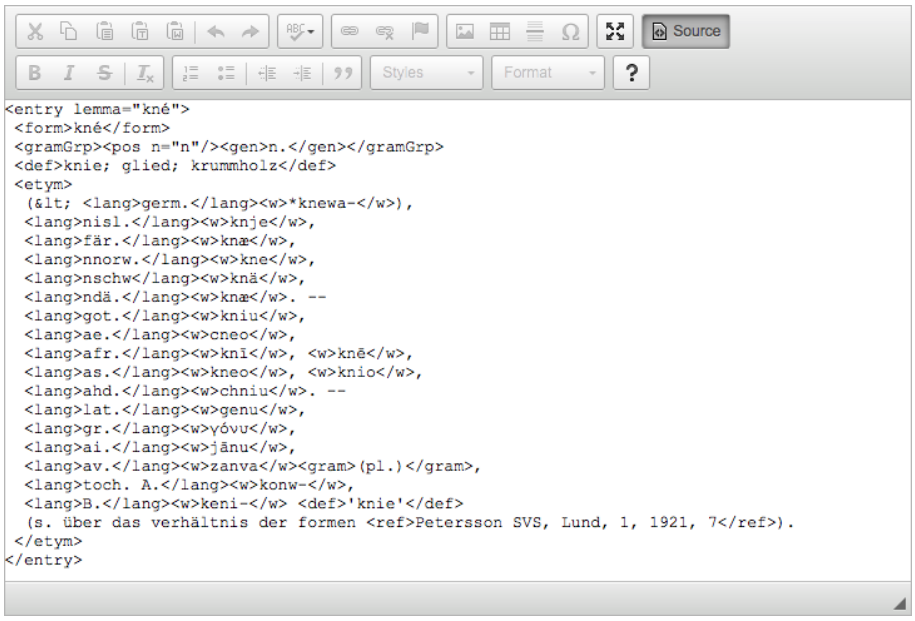
Konsortiet TEI (*Text Encoding Initiative*) blev grundlagt i 1987 med et særligt formål: at udvikle retningslinjer til opmærkning af tekster. Deres første version af *Guidelines for Electronic Text Encoding and Interchange* udkom i 1994.[1]

TEIs første *Guidelines* beskrev, hvordan forskellige typer tekster kunne opmærkes i en SGML-standard, blandt andet ordbøger. Den nuværende version – P5 – er opdateret fra SGML til XML og inkluderer flere tags til opmærkningen af ordbøger, bl.a. tagget <etym>, som indeholder alle etymologiske oplysninger og kan sammenlignes med Amsler & Tompas <E>-tag. Men når det handler om etymologier, siger TEI P5 tydeligt:

> Etymologies may contain highly structured lists of words in an order indicating their descent from each other, but often also include related words and forms outside the direct line of descent, for comparison. Not infrequently, etymologies include commentary of various sorts, and can grow into short (or long!) essays with prose-like structure. This variation in structure makes it impracticable to define tags which capture the entire intellectual structure of the etymology or record the precise interrelation of all the words mentioned. (TEI P5:9.3.4 ”Etymological Information”).

[1] Disse retningslinjer beskriver den tredje version af TEI’s opmærkningsstandard, dvs. P3. Se endvidere Ide & Sperberg-McQueen (1995).

Som TEIs retningslinjer ser ud nu, er det muligt at opmærke for eksempel *kné* fra de Vries (1961) som i Figur 3. En af grundene til at anvende TEI P5 er, at mange informationer kan opmærkes således, at den trykte ordbogs udseende kan bibeholdes. Det er for eksempel muligt at angive ordet *kné* som substantiv i tagget <pos> (*part of speech*), selvom dette ikke vises i de Vries' trykte version. Men de etymologiske informationer kan kun opmærkes ustruktureret; det er ikke muligt at gruppere de nordiske former som efterkommere eller de germanske kognater som tættere beslægtet.



Figur 3: Opslagsordet *kné* ifølge TEI P5.

2. Et nyt forslag

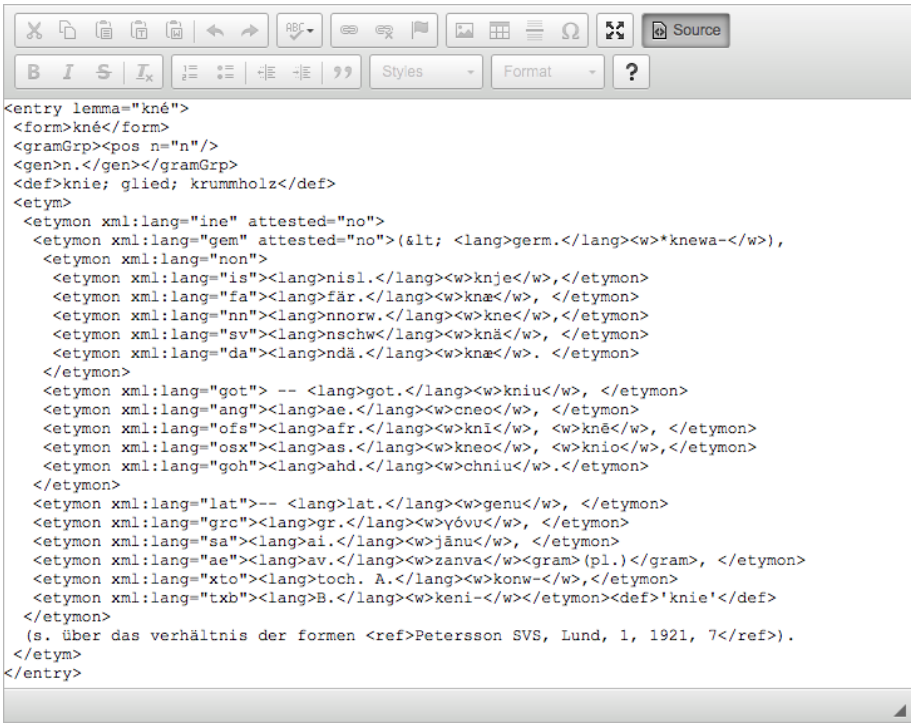
En fordel ved at bruge XML til opmærkning af etymologiske informationer er muligheden for *nesting*, dvs. at et tag kan indeholde et andet tag af samme type. Som løsning på problemet med stamtræer og digitale etymologiske ordbøger vil jeg foreslå et nyt tag, der kan indgå i TEIs retningslinjer. Tagget <etymon>, der bør placeres inde i TEIs <etym>-tag, vil kunne nestes rekursivt og indeholde tags med en kognats informationer, herunder om dets efterkommere.

Hvis vi for eksempel vil opmærke etymologien af engelsk *apple*, kan vi danne et tag <etymon>, der indeholder <w>apple</w>. Dette <etymon>-tag placeres i et andet <etymon>-tag, der indeholder den middelengelske form <w>appel</w>. Endelig anbringes dette <etymon>-tag i et tredje tag, nemlig det, der indeholder den oldengelske form <w>æppel</w>. Informationer om sprogene kan opmærkes som attributter i <etymon>-tagget, dvs. <etymon xml:lang=ang> til den oldengelske form.

Det vil også være vigtigt at kunne angive forskelle mellem former, der er belagt i skriftsproget, for eksempel oldnordisk *kné*, og rekonstruerede urformer som fællesgermansk **knewa*-. Dette kan gøres ved hjælp af et nyt attribut til <etymon>-tagget, nemlig *@attested*. På den måde vil **knewa*- kunne opmærkes i <etymon xml:lang=gem attested=no>.

Problemet med at gruppere for eksempel alle indoeuropæiske kognater kan løses ved hjælp af et tag <etymon xml:lang=ine attested=no>, som ikke indeholder et <w>-tag, fordi der ikke er angivet nogen urindoeuropæisk form hos de Vries.

Figur 4 viser, hvordan alle kognaterne placeres i indlejrede <etymon>-tags. Det yderste tag står for indoeuropæisk, som indeholder samtlige kognater, fordi de alle er indoeuropæiske sprog. Det næste tag er germansk, som indeholder den urgermanske form og dens efterkommere, de germanske kognater. Dernæst kommer de nordiske sprog, som er placeret i et <etymon>-tag, der står for oldnordisk. Til slut kommer de øvrige germanske sprog (inden for det germanske <etymon>-tag) og indoeuropæiske sprog (inden for det indoeuropæiske <etymon>-tag).




```
<entry lemma="kné">
 <form>kné</form>
 <gramGrp><pos n="n"/>
 <gen>n.</gen></gramGrp>
 <def>knie; glied; krummholz</def>
 <etym>
  <etymon xml:lang="ine" attested="no">
   <etymon xml:lang="gem" attested="no">(&lt; <lang>germ.</lang><w>*knewa-</w>),
    <etymon xml:lang="non">
     <etymon xml:lang="is"><lang>nisl.</lang><w>knje</w>,</etymon>
     <etymon xml:lang="fa"><lang>fär.</lang><w>knæ</w>, </etymon>
     <etymon xml:lang="nn"><lang>nnorw.</lang><w>kne</w>,</etymon>
     <etymon xml:lang="sv"><lang>nschw</lang><w>knä</w>, </etymon>
     <etymon xml:lang="da"><lang>ndä.</lang><w>knæ</w>. </etymon>
    </etymon>
    <etymon xml:lang="got"> -- <lang>got.</lang><w>kniu</w>, </etymon>
    <etymon xml:lang="ang"><lang>ae.</lang><w>cneo</w>, </etymon>
    <etymon xml:lang="ofs"><lang>afr.</lang><w>knī</w>, <w>knē</w>, </etymon>
    <etymon xml:lang="osx"><lang>as.</lang><w>kneo</w>, <w>knio</w>,</etymon>
    <etymon xml:lang="goh"><lang>ahd.</lang><w>chniu</w>.</etymon>
   </etymon>
   <etymon xml:lang="lat">-- <lang>lat.</lang><w>genu</w>, </etymon>
   <etymon xml:lang="grc"><lang>gr.</lang><w>γόνυ</w>, </etymon>
   <etymon xml:lang="sa"><lang>ai.</lang><w>jānu</w>, </etymon>
   <etymon xml:lang="ae"><lang>av.</lang><w>zanva</w><gram>(pl.)</gram>, </etymon>
   <etymon xml:lang="xto"><lang>toch. A.</lang><w>konw-</w>,</etymon>
   <etymon xml:lang="txb"><lang>B.</lang><w>keni-</w></etymon><def>'knie'</def>
  </etymon>
  (s. über das verhältnis der formen <ref>Petersson SVS, Lund, 1, 1921, 7</ref>).
 </etym>
</entry>
```

Figur 4: Opslagsordet *kné* i TEI samt etymologiske informationer.

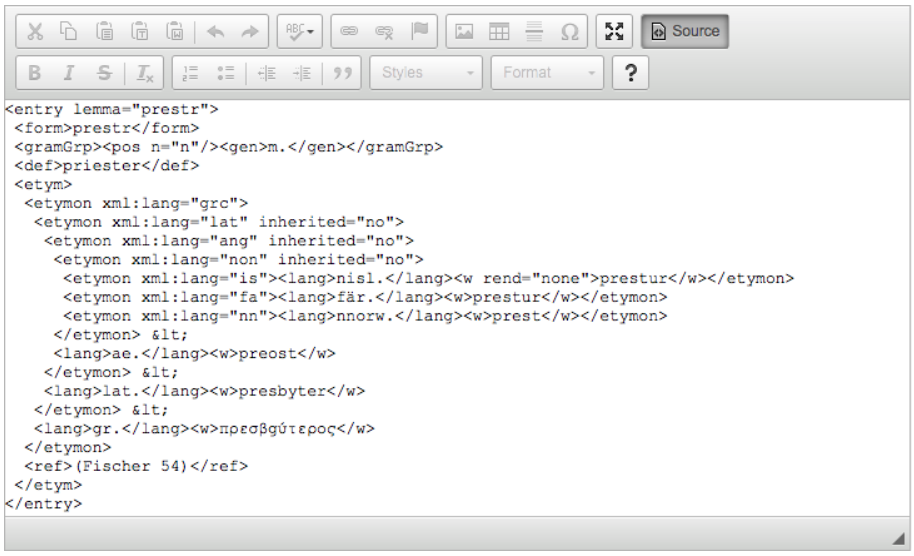
2.1. Arveord og låneord

I eksemplet *kné* er det muligt at bruge tagget <etymon> til opmærkning af etymologien, ikke mindst fordi alle sprogene er beslægtet og alle kognater nedarvet. Etymologiske ordbøger indeholder ikke kun nedarvede ord, men også et antal låneord, for eksempel *prestr* i de Vries (1961).

> **prestr** m. 'priester', nisl. fär. *prestur*, nnorw. *prest* < ae. *preost* < lat. *presbyter* < gr. πρεσβύτερος 'älter' (Fischer 54).

Ifølge Crist (2005) er der en afgørende forskel mellem arveord og låneord, som bør opmærkes på to helt forskellige måder. Vi kunne forestille os endnu et nyt tag <loan>, men det ville ikke gøre det nemmere for computeren. Men i koden behøver der ikke være så stor forskel på arveord og låneord, især hvis alle <etymon>-tags har sprogenes navne som attributter. I stedet foreslår jeg et andet attribut *@inherited*, som angiver, om et <etymon> er et arveord eller et låneord. For eksempel kunne oldnordisk *prestr* placeres i tagget <etymon xml:lang=non attested=yes inherited=no>.

Opmærkningen af *prestr* vises i Figur 5. Bemærk også, at de islandske og færøske former er identiske og derfor kun anføres én gang med den samme form (*prestur*). For at kunne inkludere den islandske form opmærkes den i et <w>-tag med attributtet *@rend=none*, dvs. at det ikke vises i den trykte bog.



Figur 5: Opslagsordet *prestr* i TEI samt etymologiske informationer.

3. Konklusion

Brugen af XML-opmærkning af etymologiske ordbøger efter TEI's retningslinjer giver mange muligheder for brugeren. Som tidligere nævnt kan man dermed opmærke informationer om både tekstens udseende og tilføje ekstra oplysninger, men der bliver tillige skabt mulighed for, at en opmærket ordbog kan bruges sammen med andre TEI-opmærkede tekster i for eksempel digitale udgaver.

Tekster behøver ikke være på papir, før de opmærkes i TEI; retningslinjerne kan også bruges til digitalt udviklede tekster. Dette er en af de vigtigste grunde til, at jeg ville udvide TEI's retningslinjer med tagget <etymon>. Som en del af mit ph.d.-projekt anvender jeg TEI P5 til opmærkning af middelalderlige håndskrifter på et bestemt oldnordisk sprog, nemlig oldgutnisk. Når jeg er færdig med at opmærke et håndskrift, hvor alle ord er tagget med et lemma, er det muligt at eksportere alle lemmata til en ekstern fil og derfra bygge en ordbog over det oldgutniske sprog. Ved anvendelse af <etymon>-tagget vil det være muligt at angive hvert ords kognater på de andre oldnordiske sprog samt en etymologi med andre germanske eller indoeuropæiske kognater.

Man kan forvente flere problemer ved anvendelse af dette tag, end jeg har beskrevet i denne artikel. For eksempel står det ikke helt klart, hvordan man kan opmærke sammensatte ord, hvor hvert led har sin egen etymologi. Men brug af et tag, som indeholder et sprogs informationer, og som kan anbringes inden i hinanden rekursivt med henblik på at beholde et stamtræs struktur, vil skabe mange nye muligheder for brugere af digitale etymologiske ordbøger.

Litteratur

Ordbøger

de Vries, Jan (1961): *Altnordisches etymologisches Wörterbuch*. Leiden: Brill.

Anden litteratur

Amsler, Robert A. & Frank Wm. Tompa (1998): An SGML-based Standard for English Monolingual Dictionaries. I: *Information in Text: Fourth Annual Conference of the UW Centre for the New Oxford English Dictionary: Proceedings of the Conference, October 26-28, 1988, Waterloo, Canada*. Waterloo, Canada, 61-79.

Crist, Sean (2005): Toward a formal markup standard for etymological data. Oakland, CA: Linguistic Society of America Annual Meeting. <www.sean-crist.com/professional/publications/crist_etym_markup.pdf> (august 2015).

Ide, Nancy M. & C. M. Sperberg-McQueen (1995): The TEI: History, Goals, and Future. I: *Text Encoding Initiative: Background and Context*. Dordrecht: Kluwer Academic Publishers.

TEI P5 = TEI Consortium (red.) (2015): *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Version 2.8.0. TEI Consortium. <www.tei-c.org/Guidelines/P5/> (august 2015).

Seán D Vrieland
ph.d.-stipendiat
Nordisk Forskningsinstitut
Københavns Universitet
Njalsgade 136
DK-2300 København S
sean.vrieland@gmail.com