PaLaFra – Entwicklung einer Annotationsumgebung für ein diachrones Korpus spätlateinischer und altfranzösischer Texte

Döhling, Lars
lars.doehling@ur.de
Universität Regensburg, Deutschland

Burghardt, Manuel
manuel.burghardt@ur.de
Universität Regensburg, Deutschland

Wolff, Christian
christian.wolff@ur.de
Universität Regensburg, Deutschland

Ziel von *PaLaFra* („Le passage du latin au français“) ist der Aufbau eines digitalen Korpus spätlateinischer und altfranzösischer Texte, das durch die Kombination von Lemmatisierung, syntaktischer und morphologischer Annotation sowie diskurspragmatischen und texttypologischen Deskriptoren komplexe Abfragestrategien ermöglicht und so eine qualitativ neuartige Nutzung der Texte bei der Rekonstruktion des lateinisch-romanischen Sprachwandels erreichen soll. Daran arbeitet ein deutsch-französisches Team der Universität Regensburg, der Universität Tübingen, der *École Normale Supérieure* in Lyon und der Universität Lille, das seit Sommer 2015 von der Deutschen Forschungsgemeinschaft (DFG) und der *Agence Nationale de Recherche* (ANR) gefördert wird. Das Projektteam ist interdisziplinär ausgerichtet und besteht aus romanischen Sprachwissenschaftlern, Computerlinguisten und Medieninformatikern. Während für den Bereich des Altfranzösischen auf das bestehende *Base de Franc#ais Me#die#val*-Korpus zurückgegriffen werden kann, so ist die Erstellung eines — was die Annotation angeht — kompatiblen Korpus spätlateinischer Texte ein wichtiges Teilziel des *PaLaFra*-Projekts.

In diesem Posterbeitrag berichten wir über Herausforderungen und Lösungsansätze bei der Erstellung einer Annotationsumgebung und eines diachronen Tagsets, das gleichermaßen in der Lage ist, die Idiosynkrasien der beiden Sprachstufen adäquat abzubilden, aber auch die diachronen Elemente im Sprachwandel einheitlich zu markieren.

Bereits für das spätlateinische Teilkorpus zeigt sich, dass es an einem standardisierten Tagset fehlt. Mindestens drei Varianten wurden in der Vergangenheit für die Annotation (spät-)lateinischer Texte entwickelt: *CoLaMer* (Selig et al. 2015), *CompHistSem* (Eger et al. 2015) und *LASLA* (Denooz 1978). Diese unterscheiden sich sowohl in den zugrunde liegenden linguistischen Konzepten als auch in ihrer Granularität. Demzufolge existiert auch kein einfaches Mapping zwischen ihnen. Für die Entwicklung eines sprachübergreifenden Tagsets in *PaLaFra* kommt erschwerend hinzu, dass die beiden Zielsprachen — Spätlatein und Altfranzösisch — trotz ihrer Verwandtschaft klare strukturellen Unterschiede aufweisen.

Zumindest für die Ebene der Wortarten (PoS, Part-of-Speech) liefert beispielsweise das Projekt *Universal Dependencies* wichtige Anhaltspunkte für ein sprachübergreifendes Tagset. Dieses Projekt hat sich die Entwicklung sprachübergreifend-kompatibler Baumbanken als Ziel gesetzt hat, die auf universellen Wortartkategorien basieren. Trotzdem bedingt die Entwicklung eines übergreifenden Tagsets oft den manuellen Vergleich von Annotationen, z.B. durch visuelle Gegenüberstellung annotierter Parallelkorpora. Unsere Recherche ergab, dass es an einem adäquaten Werkzeug für diese Aufgabe mangelt. Einerseits gibt es unzählige Annotationswerkzeuge, welche auf die Darstellung nur eines Textes samt Annotationen fokussieren (Burghardt 2014, Neves and Leser 2014). Auf der anderen Seite gibt es Alignierwerkzeuge, die auf die parallele Darstellung von Texten spezialisiert sind, aber dabei Annotation meist ignorieren, z.B. *LF Aligner* , *Moses* oder *ParaVoz* . Um diese Lücke zu schließen, haben wir auf der Basis von *InterText* — einem im Webbrowser zu bedienenden Alignierwerkzeug (Vondricka 2014) — ein Vergleichswerkzeug für annotierte Parallelkorpora entwickelt. Unsere Erweiterung unterstützt sowohl die Hervorhebung zueinander kompatibler (PoS-)Tags als auch die flexible Darstellung von Lemmata und morpho-syntaktischen Annotationen ( ). Die dafür nötigen Informationen werden beim Import aus den TEI-XML-Daten extrahiert und mit Hilfe von JavaScript dynamisch visualisiert.

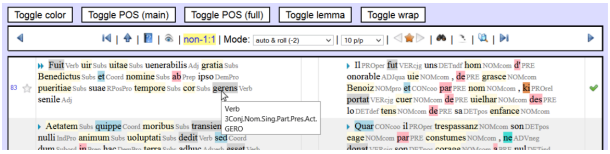

Abbildung : Das Bildschirmfoto zeigt die modifizierte InterText-Ansicht, erkennbar oben an der zusätzlichen Schalterleiste. Links ist die lateinische „Vita Benedicti“ (Vogüe und Antin 1979) zu sehen, annotiert mit dem LASLA Tagset (Denooz 1978), rechts das französische Gegenstück „Vie de saint Benoit“ (Foerster 1876), annotiert mit dem Cattex Tagset (Guillot et al. 2010). Aktuell ist sowohl die Hervorhebung kompatibler PoS-Tags („Toogle color“) als auch die Anzeige der vollständigen PoS-Annotation („Toogle POS (full)“) aktiviert.

Neben der eigentlichen Datenaufbereitung ist auch die Optimierung des Annotationsworkflows mit geeigneten Werkzeugen im Sinne verbesserter *User Experience* ein wesentliches Projektziel (*tool science*, Wolff 2015).

Die Entwicklung des spätlateinisch-altfranzösischen Tagsets wird im Projekt — auch mit Hilfe unseres modifizierten *InterText*-Tools — vorangetrieben. In unserem Posterbeitrag erläutern wir das Vorgehen und präsentieren erste Ergebnisse.

Fußnoten

1. http://www.palafra.org/
2. http://bfm.ens-lyon.fr/
3. http://universaldependencies.org/
4. http://sourceforge.net/projects/aligner/
5. http://www.statmt.org/moses/
6. https://bitbucket.org/rvwfels/paravoz2
7. http://wanthalf.saga.cz/intertext

Bibliographie

Burghardt, Manuel (2014): „Engineering annotation usability - Toward usability patterns for linguistic annotation tools“. Diss. Phil., Universität Regensburg, Institut für Information und Medien, Sprache und Kultur, urn:nbn:de:bvb:355-epub-307682.

Denooz, Joseph (1978): „L’ordinateur et le latin, techniques et methods“, in: *Revue de l’Organisation Internationale pour l’Etude des Langues Anciennes par Ordinateur* 4.

Eger, Steffen / vor der Brück, Tim / Mehler, Alexander (2015): „Lexicon-assisted tagging and lemmatization in latin: A comparison of six taggers and two lemmatization methods“, in: *LaTeCH 2015* 105.

Guillot, Céline / Prévost, Sophie / Lavrentiev, Alexei (2010): *Manuel de référence du jeu cattex09*. technical manual, UMR ICAR, CNRS/ENS-LSH. http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf

Neves, Mariana / Leser, Ulf (2014): „A survey on annotation tools for the biomedical literature“, in: *Briefings in bioinformatics* 15 (2): 327–340.

Selig, Maria / Eufe, Rembert / Linzmeier, Laura (2015): *CoLaMer* (corpus du latin mérovingien). (im Erscheinen).

Vondricka, Pavel (2014): „Aligning parallel texts with intertext“, in: *Proceedings of LREC 2014*.

Vogüé, Adalbert de / Antin, Paul (1979): *GREGOIRE LE GRAND, Dialogues* II. Cambridge University Press.

Von Foerster, Wendelin (1876): *Li Dialoge Gregoire lo Pape. Altfranzösische Uebersetzung des XII. Jahrhunderts der Dialogen des Papstes Gregor, mit dem lateinischen Original, einem Anhang: Sermo de Sapientia und Moralium in Iob Fragmenta, einer grammatischen Einleitung, erklärenden Anmerkungen und einem Glossar*, première partie: Textes. Paris: Champion.

Wolff, Christian (2015): „The case for teaching ‚tool science‘. Taking software engineering and software engineering education beyond the confinements of traditional software development contexts“, in: *Global Engineering Education Conference (EDUCON), 2015 IEEE* 932–938 10.1109/EDUCON.2015.7096085