A Althochdeutsch-Datengenerator, der

Stefanie Dipper

Anlässlich der Erstellung dieses Festschriftbeitrags stellte ich mir die Frage, welches Geschenk Karin Donhauser am meisten erfreuen würde. Die Antwort scheint mir sehr einfach und naheliegend: Eine Sammlung bisher unbekannter, originaler althochdeutscher Daten, das wäre das Größte!

Mit dieser Erkenntnis vor Augen machten wir uns ans Werk. Wir, das sind meine Mitarbeiter Marcel Bollmann, Julia Krasselt, Janis Pagel, Florian Petran, Adam Roussel und Fabian Simonjetz, ohne die dieser Beitrag in dieser Form nicht möglich gewesen wäre, und ich. Mit raffinierten Methoden gelang es uns schließlich, einen Datengenerator für Althochdeutsch zu entwickeln, eine Weltneuheit! Vorbei die Zeiten, in denen jedes schon hinlänglich bekannte und vielfach untersuchte Datum für noch eine weitere Untersuchung hin- und hergewendet und unter der Lupe inspiziert wird. Ab sofort ist es möglich, jederzeit beliebige neue Texte in garantiert originalem Althochdeutsch zu erzeugen! Einige Kostproben werden gleich präsentiert.

Da wir nun schon einmal dabei waren, bastelten wir noch etwas weiter und schufen einen Neuhochdeutsch-Frühneuhochdeutsch-Übersetzer sowie einen Bairisch-Generator! Ersterer soll auch gleich zum Einsatz kommen und uns um zunächst rund 500 Jahre in die Vergangenheit versetzen. So bereiten wir uns mental und emotional auf den eigentlichen großen Sprung von mehr als 1000 Jahren vor.

Vereeret karen dogenthafter! als erſtes ſal herczlich alles guede gewunſcht werden zum hochen vaſt tage! vßer dem ſal kurcz irclerit werden, wie dat oben beſchriebene wunder werich darmede auß ſihet. alſo: man neme alles bechant altoes hoch deutze vnd ſteckt es in dat wunder werich, die alle gevolgen von drey worten gezelt. die zalen werden als mogeliche wircklicheit angeſehen. darnach wirt dat wunder werich wie vmb getreit vnd es werden gevolgen erczeug&. wen man dat wunder werich genutzet möcht, muß man nur eyne czal eingeben. es werden dan ſo vil rede gevolgen erczeug&, wie die czal vorgib. nach diſer langen rede wollen wir dat wunder werich nu by der dat ſehen:

Uuaz thaz fihu neritа , uuant er nan ſâre bi júngoron ſine , uuío er uuidar thír io uuinne / Tho quad her in alle dinemo dionoſti , daz ſiu in deſemo chunne argoſtin ; Innan d i u áfter thiu mit ruuu giruorit fuor thara / Nemet ſon imo thaz uuib : hérro , inti mihhiloſotun gót / Tho ríetun thie ginóza , ſi ſint in then hof thes heroſten biſgofo / Tho quad imo : gifih , guot iſt uns des durft , daz thie dri genenneda ein got iſt / " Sacramentum autem , quod ſuperius , ter hebit mihela uerſtannuſſida án ímo / ia iſt ſin , the iſt ſâr filu rédii , thaz er irſlúagi in thiu ſuin , inti gizumſtigu iro giuuizſcefi ni uuarun imo himila , inti mittiu tho quam ther brutigomo , inti gotes man ſo ni ereda indi ni leerda ſo ih quád , the iz in thír / Thó giang náh ther ánther , thaz ér iſt thiſu uuitua , girihhu ſia , ir ſélbon thaz inſtúantit ana lánglicha fríſt , uuíolih er ſih bihíazi , er gotes ſún guater

leyder vert ich nichtz von den gevolgen, aber ich hoffe auf fraw dogenthafter, die ſich alles begeronge volle loeſen wirt. wer ſelber dat wunder werich genutzet wil, kan diſe ſyte auffruffen vnd dort eyne czal vnd ein wort angeben vnd kan nach hertzes lueſt altoes hoch dhucze erczeugen! ge czu diſer ſyte:

https://www.linguistics.rub.de/comphist/resources/fun/

wer wiſſen wil, wie dat frue newe hoch dhucze erczeug& wirt, der les iczt weiter. alſo: man neme reden in frue newe hoch dhucze vnd ſeczet ſy in nieuwes dhucze. darauß ergeben ſich gemeynſchaf von worten mit altem vnd neue dhucze. diſe gemeynſchaf ſteckt man in ein anders wunder werich, dat regeln lerent, wie auß allet newe wirt. möcht man frue newe hoch dhucze erczeugen, ſo muß man dat wunder werich wider vmb dregen. man ſteckt nieuwes dhucze hinein vnd es kumpt altoes dhucze herauß.

So, jetzt ist es an der Zeit, wieder in die Moderne zurückzukehren. Dafür werfen wir nun den zweiten Generator an, für Bairisch. Auch Bairisch gehört bekanntermaßen, wie Althochdeutsch, zu den sogenannten ‚less-resourced languages', so dass unser Bairisch-Generator ebenfalls eine wichtige Lücke in der linguistischen Forschung füllen wird. Wir verabschieden uns also mit einem kurzen, eigens für diesen Beitrag erzeugten Textstück in modernem Bairisch.

Omeakung: De klitisiatn Personalpronoma san durch fimf Buslinien vo da WWE au wor dort bis Backlash 2008 a Umwäidzona eingricht worn. Wai Google de wertvoiste Markn vo olle zwoa aus der ois Konzertsoi vüiföitig gnuzt wird. Ois zusezliche Obsicharung hod nia oana an aifaunga kina oda sunst wo - beispuisweis nach Thomas L.

Aha, so ist das also!

Wer es etwas genauer wissen will: Für die beiden Generatoren haben wir ein gängiges Verfahren aus der Computerlinguistik angewendet. Dazu werden zunächst Trigramme (d. h. Folgen von drei Wörtern) in vorhandenen Originaltexten, den sogenannten ‚Trainingsdaten', gezählt. Als Trainingsdaten für den Althochdeutsch-Generator haben wir sämtliche althochdeutschen Texte aus dem Referenzkorpus Altdeutsch genommen (http://www.deutschdiachrondigital.de/), die Trainingsdaten für den Bairisch-Generator stammen aus der Bayrisch-Östareichischn Wikipedia (https://bar.wikipedia.org/). Das erste Trigramm, das gezählt wird, besteht dabei aus den ersten drei Wörtern des Textes, das zweite Trigramm aus den Wörtern Nr. 2–4, das dritte aus den Wörtern Nr. 3–5 etc. Anhand der Trigramm-Frequenzen wird nun berechnet, wie hoch die Wahr-

scheinlichkeit ist, dass auf ein gegebenes Wort zwei bestimmte weitere Wörter folgen und welche weiteren Wörter mit welcher Wahrscheinlichkeit auf diese Wörter folgen etc. Diese Wahrscheinlichkeiten kommen dann beim Generieren des neuen Textes zum Tragen. Auf den Punkt gebracht bedeutet das: Der Generator hat ein sehr kurzes Gedächtnis: Nach jeweils drei Wörtern hat er schon vergessen, was er zuvor produziert hat!* Die Software, die wir zum Trainieren und Generieren nutzen, stammt von Rob Dawson (http://codebox.org.uk/pages/markov-chain-in-python).

Man mag sich jetzt fragen, ob ein solches Programm überhaupt jemals zu irgendetwas nutze ist, außerhalb von Festschriftbeiträgen. Die Antwort ist: ja! Allerdings kommen solche Programme selten als reine Generatoren wie hier zum Einsatz. Meistens werden sie zum Abgleich anderer Daten genutzt, z. B. um bei einem Text mit Rechtschreibfehlern vorherzusagen, welches (ähnlich geschriebene) korrekte Wort im vorliegenden Kontext am wahrscheinlichsten wäre.

Für den Frühneuhochdeutsch-Übersetzer haben wir ‚Norma', eine selbst entwickelte Software, eingesetzt (Bollmann et al. 2012). Dazu müssen als erstes manuell Trainingsdaten erstellt werden, die sich entsprechende alte und neue Wortformen einander paarweise zuordnen. Aus diesen Paaren werden dann Ersetzungsregeln gelernt, die angeben, welche Buchstaben(sequenzen) mit welcher Wahrscheinlichkeit durch andere Buchstaben ersetzt werden, um die eine Wortform in die andere zu transformieren. Nach der Trainingsphase, in der die Regeln gelernt werden, kommen die Regeln zum Einsatz. Das Programm bekommt eine alte Wortform als Input und wendet die Ersetzungsregeln auf diese Wortform an und transformiert sie. Das Endprodukt wird gegen ein Vollformenlexikon abgeglichen.

Dieser Vorgang wird oft ‚Normalisierung' genannt. Normalerweise wird Normalisierung eingesetzt, um historische Wortformen in die entsprechende moderne Sprache zu ‚normalisieren' und so die weitere Verarbeitung zu vereinfachen. In unserem Fall haben wir die Übersetzungsrichtung umgedreht: Input war modernes Deutsch, Output Frühneuhochdeutsch. Die Trainingsdaten für den Übersetzer stammen aus dem Anselm-Korpus (https://www.linguistics.rub.de/comphist/projects/anselm), ergänzt durch einige Texte aus dem Referenzkorpus Frühneuhochdeutsch (http://www.ruhr-uni-bochum.de/wegera/ref/).

* Einige dem Referenzkorpus Altdeutsch zugrunde liegende Editionen verwenden für *uu w*. So erzeugte der Generator den ersten Satz und *wíolih* mit *w*, sonst aber *uu*. Um eine ‚authentisch' ahd. Handschrift zu erstellen, haben wir in diesem Fall nachträglich eingegriffen.