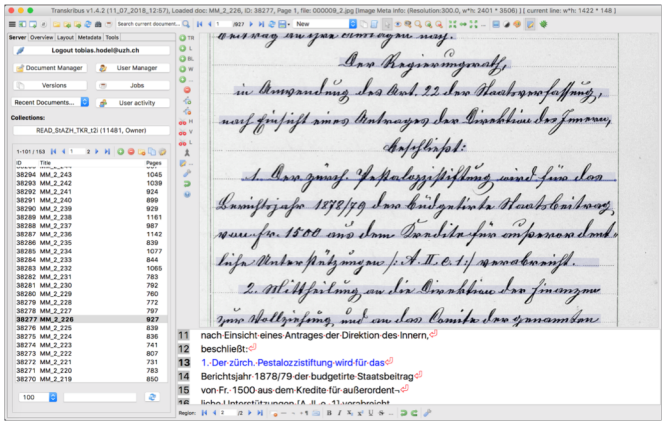
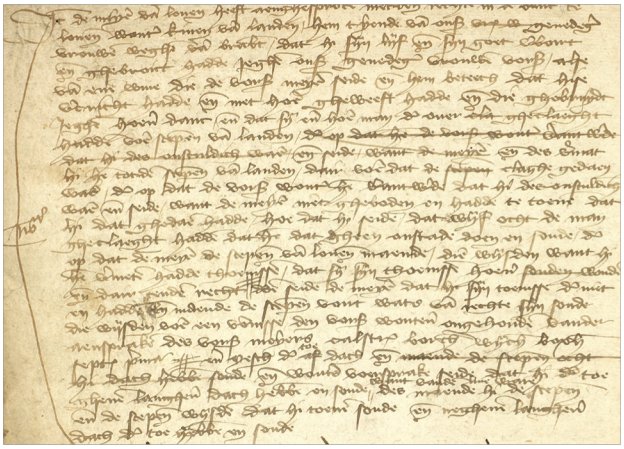
Generierung von Trainingsdaten für die Handschrifterkennung aus TEI annotierten Dokumenten

Maximilian Bryan, Tobias Hodel, J. Nathanael Philipp

Transkribus



Historische Handschriften



XML

```
<file pages="SAL7302/R°149.3" images="SAL7302/SAL7302_0297.jpg ">It(em) va(n) alsulk(er) ticht
    en(de) aenspraken als de meye(r) va(n) lov(en) gedaen heeft in/ de banc te loven(en) totden
    p(er)sone(n) hier nae ghen(oempt) te wete(n) vranc vaes jan vaes/ sijn sone ja(n) van bost
    jan va(n) wanghe jan keye ghoert va(n) rylaer pet(er)/ boccaers d...nt? scoesitte(r)? art
    vand(er) porte(n) wout(er) swilden gord hourtas/ henr(ic) van eede [gheert] =...?=
    scoesitte(r) va(n) ost gheerke(n) sijn sone gheerke(n) scoesitte(r)/ va(n) goetsenhove(n)
    gheert va(n) bost jan va(n) ...ine(n)? als dat sij met gescutte/ in brab(an)t come(n)
    soude(n) sijn en(de) hoe(r) viande te huyse en(de) te hove gesocht/ soude(n) hebbe(n) oft in
    een tav(er)ne d(aer) sij wae(re)n en(de) ald(aer) op hen ghescote(n)/ hebbe(n) souden soe
    sijn alle de vors(creven) p(er)sone nae tv(er)antw(er)de(n) dat/ sij d(aer)bi
    v(er)antw(er)den hebbe(n) vand(er) vors(creven) aensprake(n) en(de) tichte(n) des meyers/
    ghewijst quite en(de) ongehouden mo(n)te py(n)noc h(ers) alb(us) h(er)tsh(als) roel(ofs)/
    m(er)tens no(vembris) xxi</file>
<file pages="SAL7302/R°154.4" images="SAL7302/SAL7302_0307.jpg ">It(em) va(n) alsulker ticht
```

Extrahierter Text

```
It va alsulke twiste en tebatte al gheweest hebbe tussche h peten
leema in deen side en jan vand roen in dand side Soe bleve
beyde de vors ptie bleve in segghen en in vriedel yffenen dats te wete
de vors h pet in jane vande hove dross va rotsel en in lodic
ut helcht en de vors jan vand roen in lodic den rike en
in arnde de btmeke en hebbe gheloeft beide de vors ptien
vast en ghestedich te houde Soe wat de vors iiii seggen een
drechtichlec d af segghen en pnucen sele tusschen dit en
maendach naist comende op dat sijs va accorde wen En soe waen
de vors iiii segghen binne dese tide va acorde niet en worden
dat sij dan al dats voe hen come sal sijn ovgheve sele mijn he
den proeft va s gtrude als ovman ghecore tussche den vors
ptien die dach en tijd hebbe sal sijn segghe te segghe tot
s mertens misse naist comede En hebbe gheloeft de vors
ptien vast en ghestedich te houden soe wat de vors iiii
segghen op dat sijs va acorde wen tussche dit en maendach
vors ochte de vors ovman binne de tide en daghe he d toe
gheset hier af eendrechtichl segghe en pnucen sele mote
htsh oct quita
```

TEI: Standard für digitale Editionen

Maschinelle Lernverfahren benötigen Trainingsdaten um eine robuste Erkennung zu ermöglichen.

Trainingsdaten:
- bereits transkribierte graphische Dokumente
- liegen zeilenweise vor

TEI:
- strukturierte XML-Daten
- enthält Informationen über Struktur und Annotationen
- Kennzeichnung jedoch inkonsistent

Kennzeichnungen:
- zum Abschnitt zugehörige Bilddatei
- Absätze
- Zeilenumbrüche
- Personen / Orte
- Abkürzungen

TEI: Uneinheitliche Auszeichnung

Abkürzungen

XML:

```
<choice>
    <abbr>It</abbr>
    <expan>Item</expan>
</choice>
```

Klammerung:

```
It(em)
```

Zeilenumbrüche

XML:

```
<lb />
```

Doppelslash:

```
Ich bin
ein//Zeilenumbruch
```

Durchstreichungen

Textveränderungen, die eine Löschung des Textes kennzeichnen, werden in TEI standardmäßig mit <del> gekennzeichnet, Durchstreichungen bspw. mit:

```
<del rend="strikethrough">TEXT</del>
```

Manche Dokumente verwenden jedoch folgende Form: ==TEXT==

Dateinamen

Für das Training der Modelle ist es essentiell, dass für die einzelnen Bilddateien bekannt ist, welcher Text zugehörig ist.

Typisch:

```
<pb n="17" facs="Z148514201/
00000017.jpg“>
    […]
</pb>
```

Aber auch:

```
<pb n="17" facs="Z148514201/
00000017.jpg“ />
<p>
    […]
</p>
```

Dieses Projekt wurde von der Europäischen Union im Rahmen des Forschungs- und Investitionsprogramms *Horizion 2020* gefördert unter dem Grant Agreement Nr. 674943.