Nieuwe OCR voor bestaande digitale krantencollecties: de moeite waard?

Wie zoekt, die vindt? Dat is jammer genoeg niet altijd het geval bij online kranten. Gedigitaliseerde kranten worden doorzoekbaar gemaakt met behulp van *optical character recognition* (OCR), een proces waarbij afbeeldingen van de gedrukte pagina's omgezet worden naar computerleesbare tekst. Die tekst kun je vervolgens doorzoekbaar maken. Hoe beter de OCR, hoe beter de zoekresultaten. Maar het omzettingsproces is zelden foutloos. Factoren zoals de complexiteit van de lay-out, de drukkwaliteit, de conditie van het papier, de beeldkwaliteit en de gebruikte software kunnen het resultaat van OCR negatief beïnvloeden.

TEKST **Sophia Rochmes, Montaine Denys en David Coppoolse, Vlaamse Erfgoedbibliotheken**

NIEUWE TIJDINGEN

Het driejarige project *Nieuwe Tijdingen* werkt met een subsidie van de Vlaamse overheid aan een programma voor de digitalisering, duurzame bewaring en ontsluiting van het Vlaamse krantenerfgoed. Het verbeteren van de toegang tot bestaande digitale krantencollecties is daarbij een belangrijk aandachtspunt. Eén aspect is de doorzoekbaarheid. Als we de OCR kunnen verbeteren, vergroten we de vangst en de betrouwbaarheid van de resultaten bij het zoeken in deze collecties.

In de laatste jaren is er met behulp van machinelearningtechnologie veel vooruitgang geboekt bij de ontwikkeling van OCR. Het kan dus interessant zijn om de bestaande OCR van digitale krantencollecties te vervangen door nieuwe. Om te weten of dat ook echt nut heeft, moet je bepalen hoe groot het verbeterpotentieel werkelijk is.

Hoe goed of slecht is het precies gesteld met de OCR-kwaliteit van gedigitaliseerde Vlaamse krantencollecties? En hoe groot is de verbetering die we mogen verwachten van hedendaagse OCR-technologie? Om deze vragen te beantwoorden, onderzocht de vzw Vlaamse Erfgoedbibliotheken samen met meemoo, drie andere expertisepartners en tien beheerders van digitale krantencollecties de OCR van het reeds gedigitaliseerde Vlaamse krantenerfgoed.

PARTNERS

Expertisepartners:

- meemoo
- Staatsbibliothek zu Berlin
- KB Lab
- Odoma

Digitale krantencollecties van:

- Amsab–Instituut voor Sociale Geschiedenis
- Erfgoedbibliotheek Hendrik Conscience
- Erfgoedcel Waasland
- KADOC-KU Leuven
- Liberas
- Nieuws van de Groote Oorlog
- Openbare Bibliotheek Brugge
- Stadsarchief Kortrijk
- Stuifzand
- Zuidwest

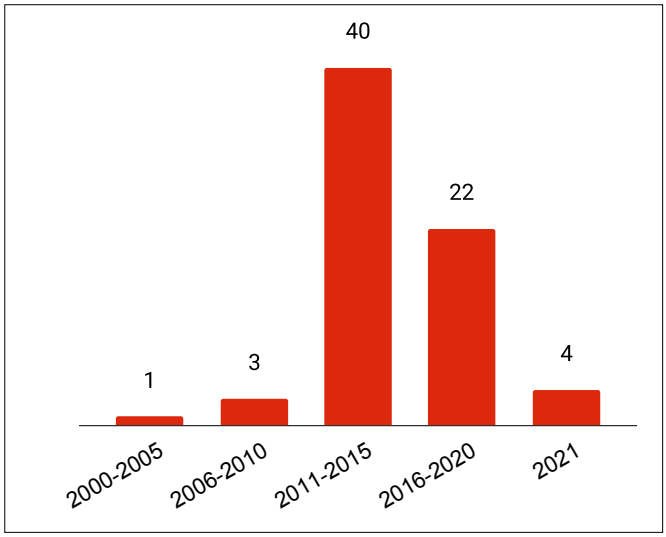
TESTDATA

Voor dat onderzoek hadden we om te beginnen testdata nodig. We selecteerden zeventig krantenpagina's uit de digitale collecties van de partnerorganisaties. Ze vormen een dwarsdoorsnede van de historische kranten die in Vlaanderen onder de scanner gingen. We kozen bewust voor variëteit qua datum van uitgave, het aantal kolommen, de taal van de inhoud, de drager (papier of microfilm), het jaar van digitalisering en het formaat van de beeld- en OCR-bestanden. Die diversiteit moest ervoor zorgen dat de resultaten van het onderzoek breed toepasbaar zouden zijn.

OCR-kwaliteit kun je niet meten zonder een foutvrije versie van de tekst. Die foutvrije versie kun je zien als het ideale OCR-resultaat, ook wel de '*ground-truth*' genoemd. Je kunt de kwaliteit van OCR meten door die te vergelijken met de *ground-truth*: elke afwijking is in principe een fout. Om te bepalen of betere OCR mogelijk is, zet je eerst de oorspronkelijke OCR af tegen de *ground-truth*. Daarna doe je hetzelfde met de nieuw gegenereerde OCR. Als de nieuwe OCR significant beter overeenkomt, is er ruimte voor verbetering.

OCR-kwaliteit gaat niet alleen over de tekst, maar ook over de lay-out. De opmaak van kranten is complex en de kwaliteit van de lay-outherkenning kan een grote impact hebben op de digitale bruikbaarheid. Bij lay-outherkenning wordt de pagina gesegmenteerd in tekstregio's en afbeeldingen. Ook de leesvolgorde wordt bepaald. Om de kwaliteit van de segmentering te meten, heb je eveneens een *ground-truth* nodig.

Het maken van *ground-truth* kost veel tijd, omdat het grotendeels handwerk is. Gelukkig zijn er wel hulpmiddelen beschikbaar. We maakten gebruik van Transkribus, een online platform voor de transcriptie en automatische tekstherkenning van historische documenten. Na het opladen van de bestanden begonnen we met de segmentering van de pagina's. We duidden de tekstblokken, groepering en leesvolgorde manueel aan. Daarna werden de tekstregels automatisch gedetecteerd door Transkribus, gevolgd door een minutieuze handmatige correctie. Vrijwilligers en collega's van de partnerorganisaties gingen vervolgens in Transkribus aan de slag met het maken van foutloze transcripties. Na een uitgebreide kwaliteitscontrole converteerden we de transcripties en de segmentering naar PAGE-XML, een standaardformaat voor het bewaren en uitwisselen van dergelijke gegevens. Met deze referentieset konden we starten met gedetailleerde kwaliteitsanalyses van de oorspronkelijke en nieuwe OCR van de zeventig pagina's.



Figuur 1: Spreiding van de pagina's over het jaar van digitalisering. De meeste pagina's zijn vrij recent gedigitaliseerd, maar ondertussen is de OCR-technologie nog sterk geëvolueerd.

Figuur 2: Segmentering en transcriptie van een krant in Transkribus. Collectie: Heemkundige Kring Meerhouts Patrimonium via Stuifzand.

Behalve met *ground-truth* kun je de OCR-kwaliteit meten door te kijken of de herkende woorden voorkomen in actuele of historische woordenboeken. Dat is minder precies, maar het voordeel is dat je zonder al te veel voorbereiding een groot aantal pagina's kunt analyseren, bijvoorbeeld om de OCR-kwaliteit van een specifieke collectie door te lichten. Voor ons project stelden we een dataset samen van duizend willekeurige pagina's uit de digitale krantencollecties van Openbare Bibliotheek Brugge en de collectie *Nieuws van de Groote Oorlog*. Met deze data wilden we nagaan hoe zinvol een woordenboekgebaseerde kwaliteitsanalyse is.

EVALUATIES

Voor de analyses werkten we samen met het Zwitserse bedrijf Odoma. Het bedrijf ontwikkelt machinelearningtechnologie die toegepast kan worden op teksten, en beschikt over hoogstaande OCR-expertise. Odoma evalueerde de kwaliteit van de tekstherkenning en de segmentering van de oorspronkelijke en de nieuwe OCR.

Nieuwe OCR

We maakten om te beginnen nieuwe OCR voor zowel de zeventig pagina's met *ground-truth* als de duizend andere, gebruikmakend van recente versies van zeven OCR-programma's: ABBYY FineReader, Amazon Textract, Google Document AI, Kofax OmniPage, Microsoft Azure Computer Vision, Tesseract-Eynollah en Transkribus. Die nieuwe OCR diende als basis om de kwaliteit van deze software te meten en om het verbeterpotentieel ten opzichte van de oorspronkelijke OCR te bepalen.

Tekstherkenning

De evaluatie van de tekstkwaliteit gebeurde vanuit verschillende invalshoeken. Twee prestatie-indicatoren bleken erg nuttig te zijn. De *Word Error Rate – Unordered* (WER-U) vergelijkt de verschijningsfrequentie van woorden in de *ground-truth* met die in de OCR, zonder rekening te houden met volgorde. De indicator meet het foutenpercentage: hoe lager de score, des te beter de OCR.

De tweede belangrijke prestatie-indicator is *Dictionary Lookup – Weighted* (Dictionary-W). Daarbij worden de woorden in de OCR vergeleken met woorden uit geschikte actuele of historische woordenboeken, rekening houdend met woordfrequentie. Hoe frequenter een woord voorkomt, hoe belangrijker het is dat het teruggevonden wordt in het woordenboek. Deze indicator meet het percentage gevonden woorden: een hogere score wijst op betere OCR.

Wat kunnen we concluderen uit de analyses van de tekstkwaliteit?

(1) Google Document AI is de software die allround het best presteert. Tesseract-Eynollah, Transkribus en Microsoft Azure hebben elk hun sterke kanten en zijn goede alternatieven.
Google Document AI scoort consequent als een van de beste bij de hoofdindicatoren en heeft geen opvallende zwaktes, ook niet bij secundaire indicatoren. Tesseract-Eynollah geeft uitstekende WER-U-resultaten. Bij bepaalde secundaire indicatoren scoort Tesseract-Eynollah echter slecht. Transkribus presteert dan weer zeer goed bij de secundaire indicatoren en minder op de hoofdindicatoren. Microsoft Azure presteert het best bij indicatoren die met woordenboeken werken.

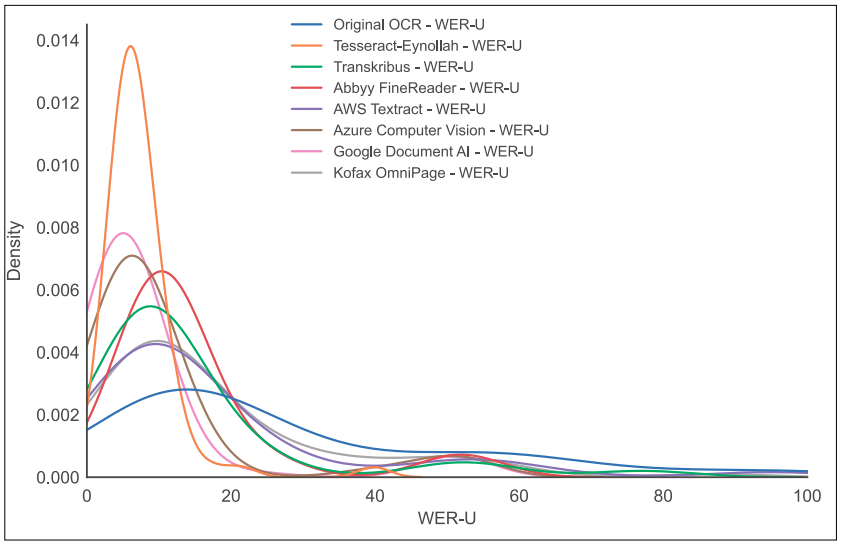
(2) De best presterende OCR-programma's hebben een verbeterpotentieel tot zeventig procent.
Alle programma's leveren gemiddeld (maar niet bij alle pagina's) betere resultaten op dan de oorspronkelijke OCR. Bij de best presterende programma's daalt de gemiddelde WER-U van 26 naar onder de 10 procent. Dat vertaalt zich in maar liefst zeventig procent minder fouten.
Figuur 3 laat zien dat de oorspronkelijke OCR (blauwe lijn) doorgaans een slechtere WER-U heeft. Alle OCR-programma's genereren veel pagina's met een lage WER-U, waarbij Tesseract-Eynollah (oranje) veruit het best presteert. Hoe goed is goed genoeg? Dat hangt af van wat je met de OCR wilt doen, maar voor het doorzoekbaar maken van historische kranten is een WER-U onder de tien procent een haalbaar streven. Raak je onder de vijf procent, dan mag je dat, gezien de complexiteit van het materiaal, welhaast perfect noemen.
Ook bij Dictionary-W (figuur 4) presteren alle programma's beter dan de oorspronkelijke OCR (blauwe lijn), met OmniPage (grijs) als achterblijver. Tesseract-Eynollah (oranje) scoort opnieuw heel goed, maar niet het best. Als negentig procent of meer van de woorden teruggevonden wordt in een geschikte woordenlijst, mag dat als heel goed beschouwd worden.

(3) De OCR van de Nederlandstalige kranten valt beter uit dan die van de Franstalige.
De puntenwolken in figuur 5 tonen de WER-U en Dictionary-W per OCR-programma. Elk punt staat voor een pagina. Idealiter bevinden alle punten zich in de linkerbovenhoek, met een lage waarde voor WER-U en een hoge voor Dictionary-W. Bij de oorspronkelijke OCR zijn de punten het meest gespreid. Bij Google Document AI, Tesseract-Eynollah en Azure liggen ze sterk linksboven. Ook andere programma's hebben een optimalere concentratie van de punten dan de oorspronkelijke OCR, opnieuw met OmniPage als minste.
Qua taal zien we een duidelijke trend: de Dictionary-W-waarde is in alle gevallen beduidend beter voor Nederlandstalige kranten (blauw) dan voor Franstalige (oranje). Misschien hebben de OCR-programma's het moeilijk met speciale tekens (zoals accenten), die meer voorkomen in het Frans, maar waarschijnlijk komt het door de evaluatiemethode. Die maakt voor het Nederlands gebruik van een historische woordenlijst die de periode van 1550 tot 1970 dekt, terwijl voor het Frans alleen een woordenboek voor het einde van de negentiende eeuw beschikbaar was. Het was dus minder waarschijnlijk dat een correct herkend Frans woord gevonden werd in de woordenlijst. De WER-U-scores zijn voor beide talen gelijklopend, wat erop wijst dat de tekenherkenning dezelfde kwaliteit heeft.

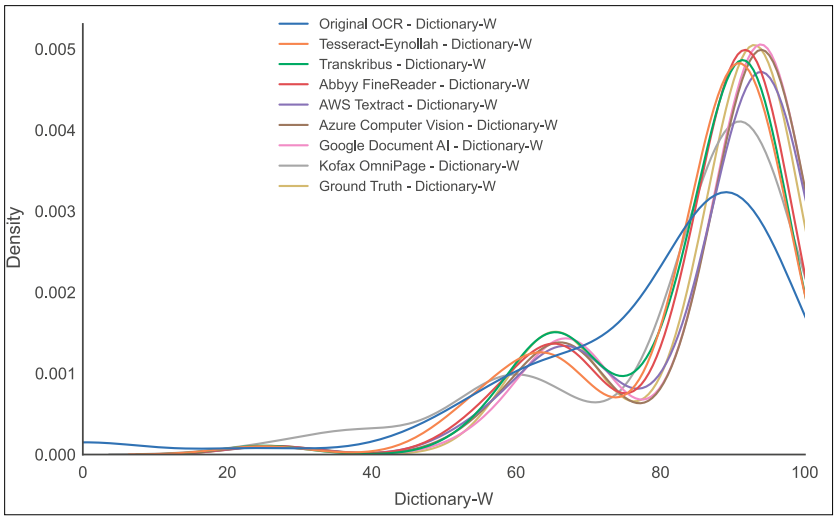
(4) De resultaten voor de testdata zonder ground-truth komen deels overeen.
Zonder *ground-truth* zijn de evaluatiemethodes beperkt. De dataset van duizend pagina's zonder *ground-truth* kon wel geanalyseerd worden met woordenlijsten als toetssteen. Ook hier valt de Dictionary-W-score voor de Franstalige kranten lager uit dan voor de Nederlandstalige kranten, gemiddeld dertig procent (figuur 6).
Google Document AI en Microsoft Azure komen opnieuw als sterkste uit de bus. Tesseract-Eynollah, Transkribus en OmniPage geven bij deze Dictionary-W-evaluatie slechtere resultaten dan bij de kleine dataset met *ground-truth* (cf. figuur 4).

		Dutch	French	Overall
Original OCR	Mean	86,1	58,9	79,1
	Median	87	61,6	84,4
Tesseract-Eynollah	Mean	82,9	50,6	74,6
	Median	83,2	48,5	79,6
Transkribus	Mean	78,6	49,9	71,3
	Median	78,2	50,2	75,2
Abbyy	Mean	85,3	54,4	77,9
	Median	87,4	59,1	84
AWS	Mean	90,2	58,5	82
	Median	90,5	60	88,3
Azure	Mean	91,5	59,6	83,3
	Median	91,9	61,8	90,1
Google	Mean	91,4	60,2	83,4
	Median	91,9	63	90
OmniPage	Mean	79,1	40	69
	Median	79,1	36,5	73,4

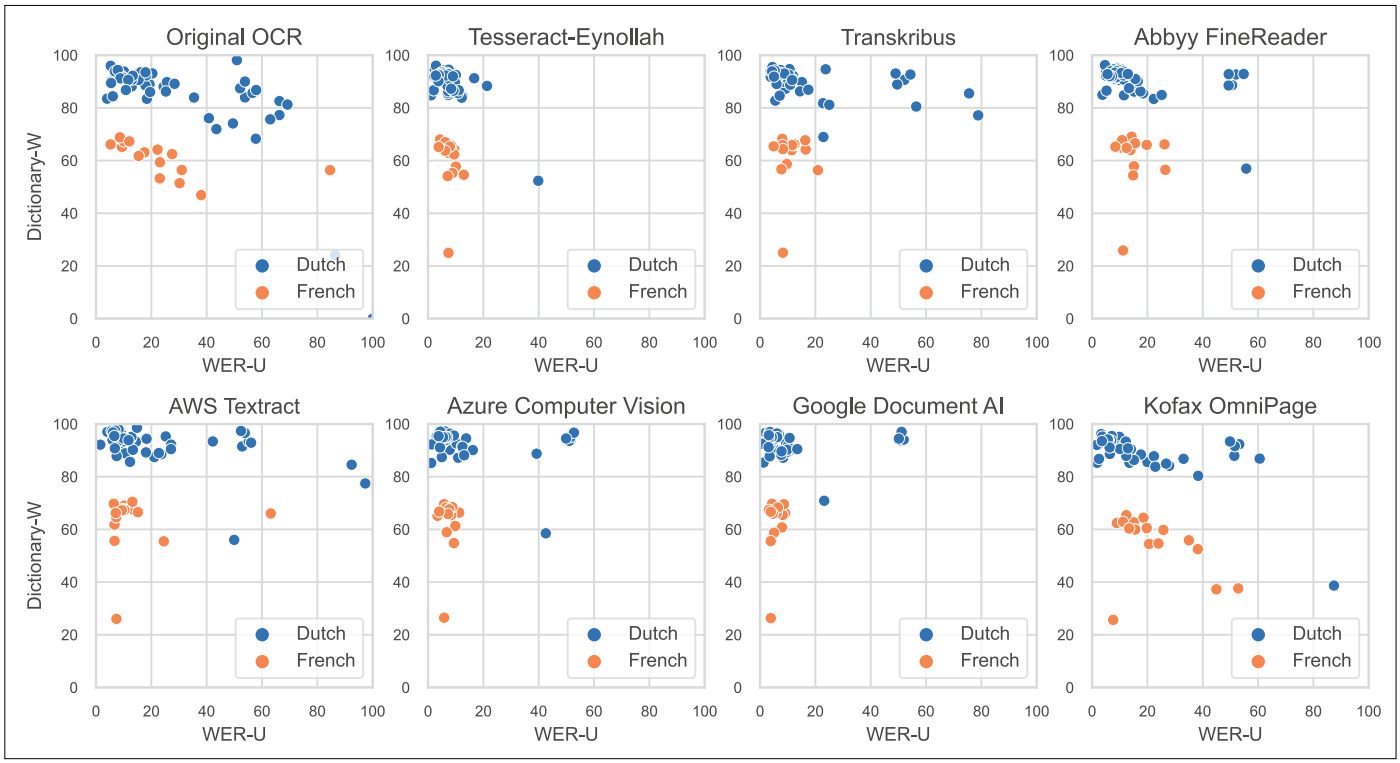
Figuur 6: Dictionary-W-scores voor de grote dataset zonder ground-truth, per programma en taal. De beste scores per kolom in groen, de slechtste in rood. Bron: Odoma.



Figuur 3: WER-U-scores van de oorspronkelijke OCR en de zeven OCR-programma's. Hoe lager de score, hoe beter de prestatie. Grafiek: Odoma.



Figuur 4: Dictionary-W-scores. Hoe hoger de score, hoe beter de prestatie. Grafiek: Odoma.



Figuur 5: Correlatie tussen de Dictionary-W- en de WER-U-scores. Optimale scores liggen in de linkerbovenhoek. Grafiek: Odoma.

Lay-outherkenning

De leesvolgorde van de lay-outherkenning werd vergeleken met die van de *ground-truth*s. Deze prestatie-indicator kijkt alleen naar de volgorde van de herkende tekstregio's, niet naar het aantal regio's dat herkend wordt. Ziet een OCR-programma slechts de helft van de tekstregio's, maar zet hij die allemaal in de juiste volgorde? Dan krijgt hij toch een perfecte score.
Aan de hand van drie bijkomende prestatie-indicatoren werd daarom gekeken of een OCR-programma de tekstregio's correct herkent. De indicator *True detection* geeft aan in welke mate de herkende regio's overlappen met die in de *ground-truth*. Hoe hoger het percentage pixels dat overlapt, hoe beter de tekstregio's gedetecteerd zijn. De indicator *Miss/partial miss* vertelt welk percentage pixels van de tekstregio's niet herkend wordt. De derde indicator, *False detection*, meet het omgekeerde: het percentage pixels dat foutief aangemerkt wordt als tekstregio. Idealiter zijn de *Miss/partial miss*- en *False detection*-scores erg laag.
Het is verder mogelijk dat een OCR-programma de tekstregio's anders opdeelt. De indicatoren *Merge* en *Split* laten zien hoeveel van de tekstregio's door het programma samengevoegd of gesplitst zijn tegenover de *ground-truth*.

Wat kunnen we concluderen uit de analyses van de lay-outkwaliteit?

(1) Tesseract-Eynollah presteert het best op lay-outherkenning, gevolgd door Transkribus, ABBYY, Google Document AI en OmniPage.
Tesseract-Eynollah (oranje lijn in figuur 8) blijkt heel nauwkeurig te zijn, met een gemiddelde *True detection*-score van 98,1 procent. Ook bij visuele controle steken de resultaten van deze software ver uit boven die van de andere. Opvallend is dat twee programma's weinig beter presteren qua lay-outherkenning dan de oorspronkelijke OCR: Amazon Textract en Microsoft Azure.

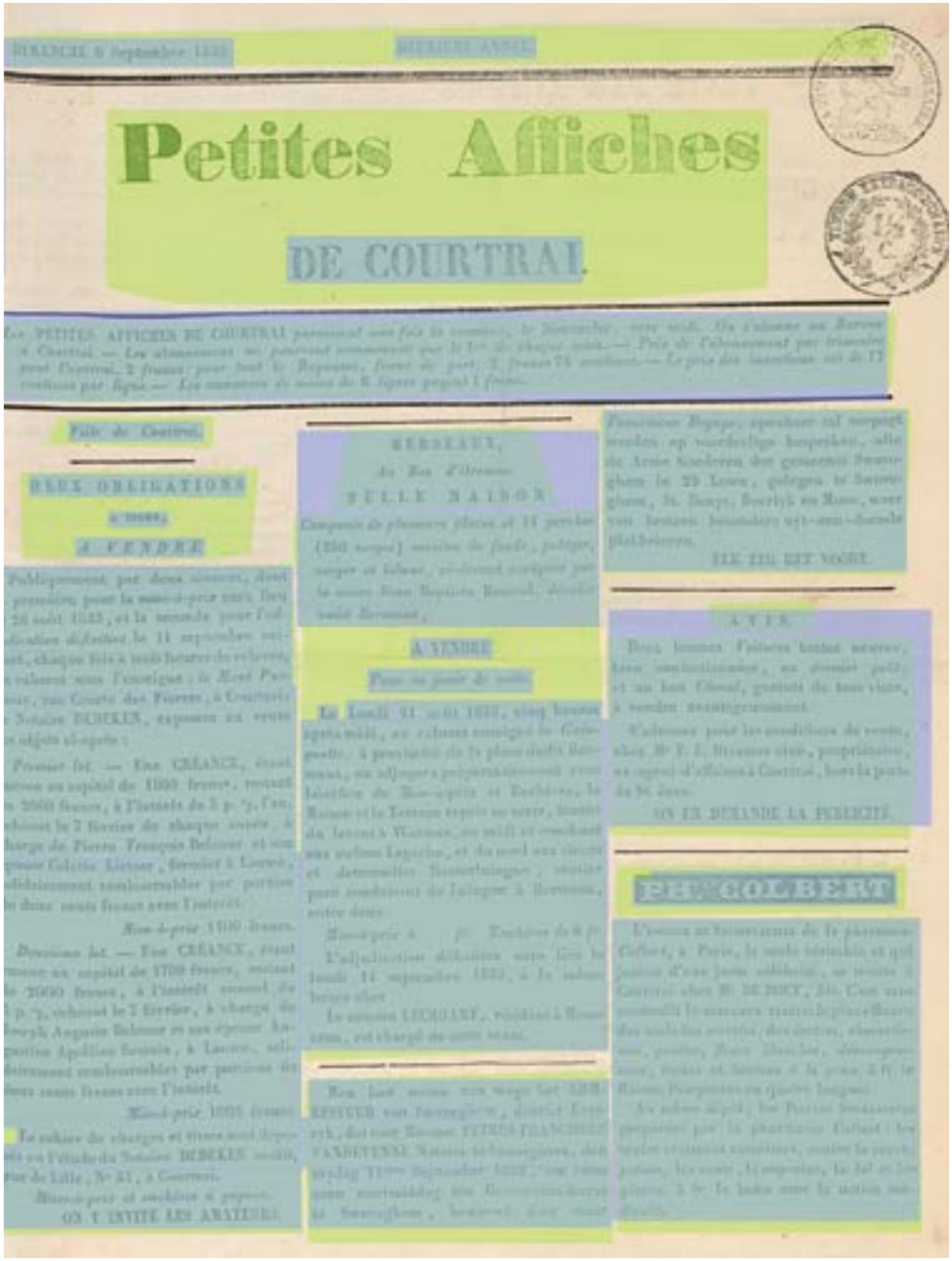
PILOOT
In 2023 voeren we binnen het project *Nieuwe Tijdingen* samen met meemoo, Stuifzand en een leverancier een piloot voor OCR-reprocessing uit. We creëren nieuwe OCR voor zo'n 150.000 krantenpagina's uit de digitale collectie van Stuifzand. Daarbij stippelen we een volledige workflow uit, van het extraheren van de beeldbestanden uit het digitale depot tot en met het herintegreren van de nieuwe OCR in het depot en de onlineontsluiting. Met deze piloot hopen we beter zicht te krijgen op de impact en de kosten van zo'n project, en op hoe je dat op grote schaal zou kunnen aanpakken.

(2) Bijna alle programma's presteren uitstekend op leesvolgorde.
In de oorspronkelijke OCR was de volgorde van de herkende tekstregio's al meer dan 99 procent correct. Bijna alle programma's scoren zeker zo hoog.

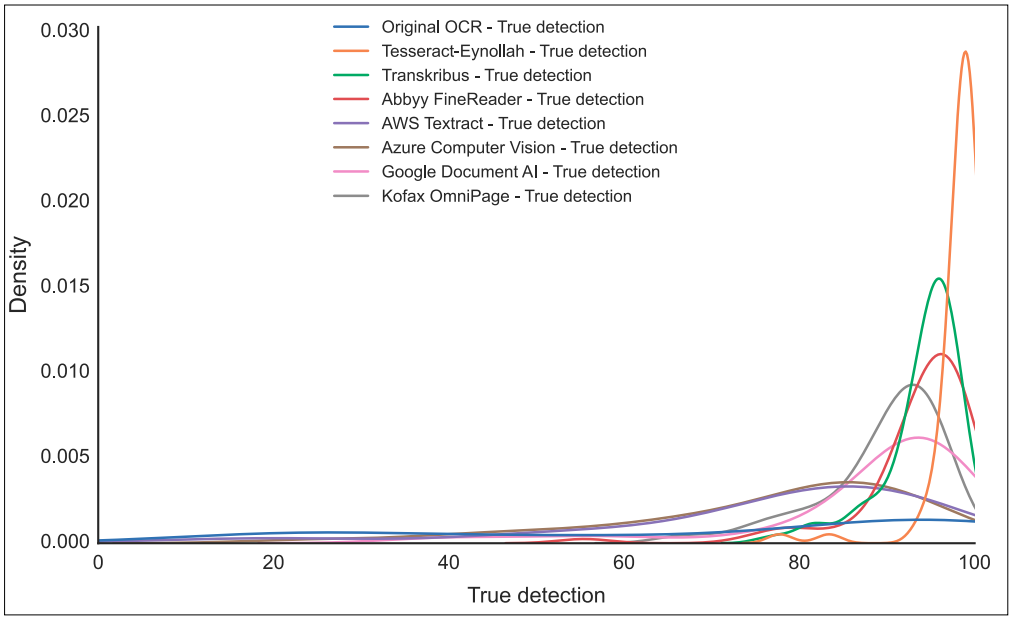
(3) De meeste programma's splitsen de tekst op in meer regio's, maar dat is niet noodzakelijk een probleem.
In principe willen we lage *Split*-scores zien, maar hoge scores leiden niet noodzakelijk tot slechtere tekstherkenning. Daarvoor maakt het niet uit over hoeveel regio's de tekst verdeeld is. Soms is er ook gewoon sprake van een andere aanpak. Bepaalde programma's splitsen de tekst op in alinea's, terwijl in de *ground-truth* alinea's binnen eenzelfde artikel aangeduid zijn als één regio.

CONCLUSIE

Is het mogelijk om de OCR van digitale krantencollecties in Vlaanderen te verbeteren? Op basis van ons onderzoek denken we van wel. Met de technologie van vandaag kun je OCR produceren die tot zeventig procent beter is dan de oorspronkelijke. En dat kan zeker een positieve impact hebben op de doorzoekbaarheid.
Welke collecties hebben er het meest baat bij? Odoma onderzocht of bepaalde aspecten van invloed zijn: jaar van uitgave, taal, collectie van herkomst, aantal kolommen, jaar van digitalisering, drager (papier of microfilm) en uitvoerformaat. Die analyse leverde geen bruikbare indicatoren op.
Er was het verschil tussen de Nederlands- en Franstalige kranten bij de Dictionary-W-*metric*, maar zoals gezegd komt dat waarschijnlijk door de beperkte woordenlijst die bij de analyses gebruikt werd. Ook viel op dat de pagina's uit één specifieke digitale krantencollectie lastig zijn voor alle OCR-programma's, behalve voor Tesseract-Eynollah. Helaas is het niet duidelijk waarom.
Er lijkt vooralsnog geen correlatie te zijn tussen de hierboven genoemde kenmerken van de collecties en het OCR-verbeterpotentieel. We kunnen dus niet zeggen of de ene collectie meer baat zal hebben bij nieuwe OCR dan de andere.
Dit onderzoek was in de eerste plaats bedoeld om inzicht te krijgen in de oorspronkelijke OCR-kwaliteit van digitale krantencollecties in Vlaanderen en het verbeterpotentieel. Maar de resultaten maken ook duidelijk welk kwaliteitsniveau je vandaag de dag mag verwachten van een OCR-programma. En ze helpen bij de selectie van een specifiek pakket binnen een digitaliseringsworkflow.
Daarbij zijn er meerdere goede keuzes. Zowel de commerciële software Google Document AI als de opensourcesoftware Tesseract-Eynollah leveren uitstekende resultaten. Andere programma's, zoals Transkribus en Microsoft Azure, presteren ook goed. De uiteindelijke keuze hangt af van meer dan alleen de OCR-kwaliteit. Hoeveel kost de tool? Welke input- en outputformaten worden ondersteund? Zijn er extra stappen nodig, zoals het verkleinen van te grote beeldbestanden? Hoe gebruiksvriendelijk is de software? Heeft jouw leverancier er ervaring mee? Allemaal zaken om rekening mee te houden.



Figuur 7: Lay-outherkenning op een krantenpagina. Donkergroen = *True detection*, lichtgroen = *Miss/ partial miss*, paars = *False detection*. Collectie: Stadsarchief Kortrijk, visualisatie: Odoma.



Figuur 8: *True detection*-scores. Hoe hoger de score, hoe beter de prestaties. Grafiek: Odoma.