TESTDAG SPRAAKTECHNOLOGIE VOOR VERSLAGGEVING VLAAMS PARLEMENT (4 MEI 2018): RAPPORT

Opzet en verloop van de testdag

In het kader van het project 'Spraaktechnologie voor verslaggeving Vlaams Parlement' binnen het Programma Innovatieve Overheidsopdrachten (PIO) vond op 4 mei 2018 in het Vlaams Parlement een testdag plaats met als doel de huidige technologische mogelijkheden van automatische transcriptie van Nederlandstalige spraak naar tekst te valideren en verder af te stemmen op de behoeften van de verslaggevende diensten van het Vlaams Parlement. De dag vormde de eindfase van de marktconsultatie ter zake. Mede op basis van de bevindingen over de resultaten die op de testdag werden afgeleverd, moet het Vlaams Parlement beslissen welke stappen, al dan niet in partnerschap met het PIO, verder zullen worden gezet naar de ontwikkeling, validering of aankoop van een innovatieve oplossing die kan beantwoorden aan de gestelde behoeften en doelstellingen inzake ondersteuning van parlementaire verslaggeving.

Voor de testdag werden drie deelnemers geselecteerd:
1. Parladium (consortium MyForce / Telecats / KU Leuven / Universiteit Gent / Radboud Universiteit Nijmegen)
2. Consortium Cedat85 / Bertin IT
3. Zoom Media

Vooraf kregen de deelnemers testmateriaal ter beschikking. Het betrof een lijst van plenaire en commissievergaderingen (september 2017 – februari 2018) waarvan een Woordelijk Verslag is gemaakt, met telkens de datum en het tijdstip én een link naar de website van het Vlaams Parlement waar zowel het videoverslag als het Woordelijk Verslag van deze vergadering kon worden teruggevonden.

Op de testdag kregen alle deelnemers drie dezelfde, voor hen onbekende audiofragmenten aangeleverd in mp3-formaat (twee van iets meer dan twintig minuten, één van iets meer dan dertig minuten), zonder vermelding van de naam van de sprekers en uiteraard zonder link naar de oorspronkelijke verslagen. Die audio-opnames moesten ze met hun spraak-naar-tekstmodel omzetten naar tekstuele verslagen. Om de resultaten te kunnen beoordelen werd enkel een vlakke tekst gevraagd. Daarnaast demonstreerde en/of beschreef elk van de verschillende deelnemers een onmiddellijk editeerbare tekstoutput, gelinkt (op woord- , segment- of zinsniveau) aan het betreffende audiobestand, waarbij via kleur- of grijswaardecodering een 'confidence level' toegekend wordt aan elk woord in de output.

Na de testdag werden de resultaten door redacteurs Woordelijk Verslag beoordeeld op basis van vooraf bepaalde en aan de deelnemers meegedeelde criteria (zie Bevindingen, 2). De diverse teksten werden hun anoniem aangeleverd.

Bevindingen

1. Algemeen

1.1. De output werd door de drie deelnemers zeer snel aangeleverd, ongeveer een uur na de start.

1.2. De kwaliteit van de geleverde teksten varieerde van redelijk bruikbaar tot ronduit onbruikbaar. Belangrijke vaststelling daarbij is evenwel dat die kwaliteit niet zozeer afhangt van de spraaktechnologie als zodanig, dan wel van de spreker zelf. De manier van spreken beïnvloedt immers vanzelfsprekend de kwaliteit en dus de bruikbaarheid van de output, zowel op woord- als op zinsniveau. Bij de (zeldzame) goede sprekers die in de geselecteerde fragmenten voorkwamen, was het resultaat behoorlijk goed.

Woordherkenning hangt onder meer af van idiosyncratische spraakkenmerken, (het gebrek aan) articulatie, mogelijke versprekingen of haperingen en regionale accenten eigen aan de sprekers. Met dergelijke accenten, die in alle mogelijke varianten in het Vlaams Parlement, en dus ook in de geselecteerde fragmenten, voorkomen, kan de spraaktechnologie duidelijk niet overweg. Heel vaak hanteren sprekers bovendien een verkeerde zinsbouw of lopen ze verloren in hun eigen, soms ellenlange, zinsconstructies, onderbreken, herhalen of verbeteren ze zichzelf of gebruiken ze stopwoorden. Zelfs indien de spraakherkenning op woordniveau feilloos zou werken, levert dat vanzelfsprekend nog steeds slechte teksten op. De omzetting van spreek- naar schrijftaal is met andere woorden nog een bijkomende, extra dimensie van spraakherkenning en artificiële intelligentie. De succesgraad zou zonder twijfel hoger liggen bij geluidsfragmenten met sprekers die duidelijk articuleren, een tekst aflezen, een minder sterk regionaal accent hebben enzovoort. Omgekeerd zou men uiteraard ook kunnen poneren dat de artificiële intelligentie op dit moment nog niet ver genoeg gevorderd is om met onze moeilijke sprekers overweg te kunnen.

1.3. Dat de beoordeling, zoals eerder bepaald en aangekondigd, gebeurde op basis van de vlakke teksten die door elk van de drie deelnemers werden aangeleverd, beïnvloedde dat oordeel in negatieve zin. Tijdens de testdag zelf demonstreerden of beschreven de verschillende deelnemers immers elk hun eigen toepassing met onmiddellijk editeerbare tekstoutput (zie hoger). Die maakt het voor de redacteurs eenvoudiger om snel naar een bepaald deel van het audiofragment te navigeren en de tekst te corrigeren dan dat op basis van een vlakke tekst, los van het audiosysteem, het geval is. Bij Cedat85 is heel dat proces bovendien geïntegreerd in een volledig workflow management systeem, dat bijvoorbeeld in het Italiaanse parlement wordt toegepast.

2. Beoordeling op basis van de verschillende criteria

Zoals beschreven in de oproep tot deelname aan de testdag, werden de resultaten door de redactie van het Woordelijk Verslag beoordeeld aan de hand van volgende criteria:

1. de tijd nodig voor de verslaggevers om elk stuk tekst om te zetten naar een 'afgewerkt verslag';
2. het type fouten of te verbeteren passages (al dan niet gerelateerd aan een welbepaalde spreker of grammaticale constructie, al dan niet 'eenvoudig' semiautomatisch te remediëren, …);
3. het percentage van accuraatheid van de sprekersherkenning;
4. de gelijkenis met het oorspronkelijke verslag.

Hierna overlopen we elk van deze criteria.

2.1. De tijd nodig voor de verslaggevers om elk stuk tekst om te zetten naar een 'afgewerkt verslag'

Ook wanneer de woorden van de spreker vrij accuraat werden weergegeven, wat overigens lang niet altijd het geval was (zie 2.2.), hadden de redacteurs nog veel werk met het aanbrengen van interpunctie en het plaatsen van hoofdletters, grammaticale aanpassingen, het schrappen van ballast zoals versprekingen of nodeloze herhalingen, het volledig herschikken van zinnen, enzovoort. Precies doordat er nog zoveel aanpassingen nodig waren, leverde het gebruik van spraak-naar-teksttechnologie slechts een veeleer beperkte tijdswinst op; bij slechte sprekers was die tijdswinst zelfs nihil. Dat de redacteurs wat minder zelf moeten tikken, wordt dan niet langer als een voordeel ervaren. Integendeel, de tijd om alles te beluisteren en de fouten in de weergave te corrigeren, telt zwaarder door. Het kost immers veel meer concentratie om een tekst te herschrijven dan om hem zelf 'from scratch' uit te tikken. Zeker bij sprekers die moeilijk uit hun woorden geraken, zijn de zinsstukken die het systeem toch heeft weten te vatten, dus slechts een zeer beperkte hulp. Zoals eerder gezegd (zie 1.3.), moet deze

conclusie echter worden genuanceerd doordat de beoordelende redacteurs geen gebruik hebben gemaakt van de onmiddellijk editeerbare tekstoutput, die elk van de drie deelnemers wel ter beschikking hebben.

De redacteurs beschouwen de resultaten die tijdens de testdag werden aangeleverd, als een stap terug in vergelijking met die van het programma Dragon NaturallySpeaking, waar ze nu al gebruik van maken, omdat er opnieuw veel meer tik- en muisklikwerk aan te pas komt, wat als meer belastend voor de gewrichten en voor nek, rug en schouders ervaren wordt.

2.2. Het type fouten of te verbeteren passages

Het soort fouten hebben we in algemene termen al in 1.2. besproken.
Hier willen we daar meer in detail op ingaan.

2.2.1. Op woordniveau

In het algemeen valt inzake kwaliteit van woordherkenning bij de verschillende deelnemers geen eenduidige lijn te trekken. Elk team heeft fragmenten waar het beter scoort dan de andere. Fouten op woordniveau kunnen zowel te maken hebben met de eigenheden van de diverse sprekers, als met het soort woorden dat moet worden weergegeven.

2.2.1.a. Eigenheden van de sprekers

- Regionale accenten en/of tussentaal

Regionale accenten en/of het gebruik van tussentaal vormen een belangrijke bron van fouten. Geen enkel van de drie teams weet daarmee om te gaan, en al helemaal niet in combinatie met het 'inslikken' van woorddelen. Enkele voorbeelden (niet exhaustief):

* Antwerps accent (in combinatie met tussentaal):
    - 'ge moet'-> gemoed
    - 'dan sluit je' -> 'dan sluiten'
    - 'een moeilijke' -> 'noem moeilijke'
    - 'gaat zeggen' -> 'gas zeggen'
    - 'als je twee toppers hebt, kun je' -> 'als getweet opgezet kunnen'
    - 'dat niet' -> 'Danny'
* West-Vlaams accent
    - 'in orde blijft' - > 'en harde beleefd'
    - 'werf gaan bekijken' -> 'Vijf Ham bekeken'
    - 'is er' -> Esther
    - 'strenge bril' -> 'Strange Brel'
    - 'kwaliteitscontrole' -> kwaliteits komt drollen'
    - 'Dank u wel' -> 'Een kwaal'
    - 'gelijk met haar bezorgdheid -> 'leek het Patershof Tate
    - 'zijn' - > 'zien'
    - 'kinderen' -> kende
    - 'beleefd geformuleerd'-> 'verliefd geformuleerd' / 'de liefde formuleert'
    - 'gans' -> 'hans'
    - 'collega's' -> 'cowley has'
* Oost-Vlaams accent (in combinatie met tussentaal)
    - 'zeg je: kijk, wij staan'-> 'zegt gevang keken wij staan'
    - 'dat je ballon' -> 'dat duwen ballon'
    - 'duidelijk' -> 'deugdelijk'
* Limburgs accent
    - 'blijven' -> 'beleven'

- Idiosyncratische spraakkenmerken

* Moeilijkheden met de uitspraak van bepaalde letters of klanken
Enkele voorbeelden:
'waar'-> 'waeg' / 'waag', 'daar' -> 'daag'
'door' -> 'dooi'
* Inslikken van woorddelen
Enkele voorbeelden:
'redelijke aanpassing' -> 'hele kampt'
'iedereen' -> in drie
Sprekers die woorddelen inslikken in combinatie met een regionaal accent, zijn helemaal een ramp om weer te geven.
* gebruik van stopwoorden zoals 'euh' (wordt soms 'aan' of hij')

2.2.1.b. Soorten woorden

- Cijfers, percentages en jaartallen

Cijfers, percentages en jaartallen worden door twee van de drie teams zelden of nooit als dusdanig herkend en dus ook niet als getal weergegeven. Bij een team daarentegen is dat nauwelijks een probleem (vb. 400 meterbaan, 0,5%), alleen een zwaar regionaal accent kan daar moeilijkheden veroorzaken. (vb. 'twee te deze' i.p.v. 120.000).

- Eigennamen, namen van instellingen en specifieke begrippen

* Ook in de weergave van eigennamen en namen van instellingen is een team over het algemeen beter dan de andere twee. Enkele voorbeelden:
Worden herkend: Bart Swings, Turnhout, Ira Vannut, Blosocentra, Wezenberg, SHM (wel niet altijd), VMSW, VVH, M-decreet, Patrick Vandelanotte
Maar: 'de Olympische Spelen in Rio' -> 'de linkse spelen in Milau', Blosocentra -> 'bloot zoals centra', Annouri -> 'aan noeri', 'Van Volcem' -> 'van vossen'

* Anderzijds worden door alle teams een (relatief groot) aantal woorden onterecht als eigennamen geïnterpreteerd.
Enkele voorbeelden:
'grote valkuil' -> 'Rode Valken' / 'rotte valken'
Zie ook bij 2.2.1.a. (regionale accenten)

- Ook moeilijke en samengestelde woorden vormen een probleem.
Enkele voorbeelden:
'multidisciplinair' en 'ondersteuningsnetwerken' worden niet herkend
'inclusie' -> 'ik Lucy'

- Vreemde woorden worden nogal eens vernederlandst of niet herkend
Enkele voorbeelden:
'centers of excellence'
-> 'centers af. Ik zal eens' of 'samen zijn examens'
-> 'Syntus als excellent' of 'recent is of Eksel dans'
-> 'centrum of excellent'

2.2.2. Zinsstructuur

De vlakke tekst die we aan de deelnemers hadden gevraagd, werd door twee van de drie teams op een onoverzichtelijke manier afgeleverd. Elke vorm van interpunctie ontbreekt, en bijgevolg staan er ook geen hoofdletters in de tekst, behalve bij eigennamen (of wat

als zodanig geïnterpreteerd wordt). Een team levert wel interpunctie af, maar niet altijd op de juiste plaats. Het maakt de tekst niettemin veel overzichtelijker.

Zinsconstructies kloppen bijna nooit. Dat heeft, zoals al in 1.2. aangehaald, vooral te maken met de slechte spreekstijl van de meeste sprekers en, bijgevolg, de bijzonder moeilijke omzetting van het gesproken woord naar begrijpelijke geschreven taal. Parladium beschikte normaliter over een 'Spreek-naar-Schrijf'-module zoals ontwikkeld voor het Nederlandse parlement (zonder aanpassingen aan de Vlaamse situatie), maar tijdens de testdag was er, om ongekende redenen, een probleem met deze module, waardoor de output ervan niet gegenereerd kon worden.

2.3. Het percentage van accuraatheid van de sprekersherkenning

Op het vlak van accuraatheid van de sprekersherkenning sprong een team er duidelijk bovenuit, met een bijna volledige sprekersherkenning (gemiddeld 90%). Dat heeft er ongetwijfeld mee te maken dat dit team, voorafgaand aan de testdag, 134 sprekersprofielen had aangemaakt, overeenkomstig de huidige samenstelling van het Vlaams Parlement en de Vlaamse Regering. De andere twee teams hadden slechts een selectie van parlementsleden uit het beschikbare testmateriaal genomen om sprekersprofielen aan te maken, met een beduidend lagere sprekersherkenning tot gevolg (gemiddeld 30%).

2.4. De gelijkenis met het oorspronkelijke verslag

Door de bevindingen die zijn weergegeven in 2.2. en 2.3. zal het niet verbazen dat de gelijkenis met het oorspronkelijke verslag (zeer) te wensen overlaat.

Conclusie

In haar huidige stand biedt de spraak-naar-teksttechnologie onvoldoende goede resultaten om nuttig bruikbaar te zijn voor de verslaggevende diensten van het Vlaams Parlement. Op dit moment levert het gebruik ervan in de context van het Vlaams Parlement geen efficiëntiewinst op, maar eerder verlies.

Er zijn dus nog belangrijke investeringen nodig in de verdere ontwikkeling van de technologie. Vraag is of het Vlaams Parlement daarbij een voortrekkersrol moet spelen en daarin op dit moment verder financieel moet investeren. Het Vlaams Parlement kiest er eerder voor de verdere evolutie van de technologie en de markt nauwgezet te blijven opvolgen en ter zake ook geregeld contact te houden met andere parlementaire assemblees, vooral dan in het Nederlandse taalgebied, om als 'early adopter' te kunnen fungeren op het moment dat de technologie voldoende maturiteit bereikt om alsnog efficiënt te kunnen worden ingezet.

Het Vlaams Parlement nodigt de betrokken onderzoeksinstellingen en marktspelers uit de videobestanden van de vergaderingen, die als open data beschikbaar zijn op zijn website, ten volle als testmateriaal te benutten, met het oog op de verdere ontwikkeling van de technologie.