Computerlinguïstiek en fraseologie

Tiberius, C.P.A.

Prof.dr. C.P.A. Tiberius

Computerlinguïstiek en fraseologie

Universiteit Leiden

Computerlinguïstiek en fraseologie

Oratie uitgesproken door

Prof.dr. C.P.A. Tiberius

bij de aanvaarding van het ambt van hoogleraar

Computerlinguïstiek

aan de Universiteit Leiden

op maandag 26 februari 2024

Universiteit
Leiden

Mevrouw de rector magnificus, zeer gewaardeerde toehoorders,

Met heel veel plezier aanvaard ik vandaag officieel mijn benoeming tot hoogleraar Computerlinguïstiek.

Computerlinguïstiek
Computerlinguïstiek is het vakgebied dat zich bezighoudt met de wetenschappelijke studie van natuurlijke taal vanuit een computationeel perspectief. Computerlinguïsten ontwikkelen computermodellen voor taalkundige verschijnselen. Enerzijds kunnen deze modellen gebruikt worden om taalkundige theorieën te toetsen. Anderzijds kunnen ze ingezet worden voor natuurlijke taalverwerking door de computer. Met name dat laatste, de ontwikkeling van computersystemen die natuurlijke taal kunnen analyseren en genereren, heeft recent een enorme vlucht genomen. Denk aan programma's die automatisch spelfouten corrigeren en geavanceerde schrijfadviezen geven, automatische vertaalsystemen zoals Google Translate, en chatbots, waarvan ChatGPT[1] waarschijnlijk de bekendste is. Dergelijke toepassingen zullen steeds meer een onderdeel worden van ons dagelijks leven.

In de begintijd, in de jaren 60 van de vorige eeuw, was de computerlinguïstiek stevig ingebed in de taalkunde. Hugo Brandt Corstius (1978), de vader van de computertaalkunde in Nederland, ontwikkelde zijn computermodellen met als belangrijkste doel het toetsen van taalkundige theorieën. Tegenwoordig, met de 'revolutie' van de Large Language Models (LLM's), de grote taalmodellen, lijkt de taalkunde nog maar weinig in de melk te brokkelen te hebben. Met grote hoeveelheden data en veel computerkracht worden indrukwekkende resultaten verkregen. Echter, zoals Emily Bender (2022) stelt:

> '[...] the more we are in awe of what this technology can supposedly do the less well positioned we are to counteract that power.'

Ik ben er dan ook van overtuigd dat voor succesvolle computerlinguïstiek, kennis van beide componenten, dus zowel taalkunde als informatica, cruciaal is.

Computerlinguïstiek vormt het eerste luik van mijn onderzoek. Lexicografie, het tweede. Ze komen samen in de fraseologie, de studie van woordcombinaties. Woordcombinaties vormen een grote uitdaging voor zowel de computerlinguïstiek als de lexicografie.

Lexicografie
Lexicografie is het vakgebied dat zich bezighoudt met het maken van woordenboeken. Woordenboeken - eentalig, tweetalig of meertalig - bevatten traditioneel de meest solide en nauwkeurige kennis over woorden en hun betekenissen. Daarnaast zijn woordenboeken ook belangrijk als culturele en historische artefacten. Zij laten zien hoe de woordenschat zich ontwikkelt en welke woorden belangrijk zijn in een bepaalde periode in de geschiedenis. Voor 2020 hadden we al wel het woord *corona*, maar we kenden het vooral in de andere betekenis 'buitenste atmosfeer van de zon'. Ook zijn er door corona veel nieuwe woorden bij gekomen die voornamelijk in de context van de wereld van 2020 en 2021 gezien moeten worden. Een zin als *Is een terrasbubbel even groot als de buitenbubbel van tien, of blijft terrassen beperkt tot uw knuffelcontact plus één?*[2] heeft betekenis in een wereld met lockdowns, maar is daarbuiten veel lastiger te interpreteren.

Ook binnen de lexicografie kunnen een aantal belangrijke verschuivingen waargenomen worden.

Van papieren woordenboek naar online woordenboek
Het zal u niet ontgaan zijn dat de woordenboekenmarkt al jaren onder druk staat. Veel uitgeverijen zijn verdwenen en de enige keer dat jongeren tegenwoordig gebruikmaken van een woordenboek is waarschijnlijk tijdens de examens op de middelbare school omdat het gebruik van een papieren woordenboek dan is toegestaan. Hoe nuttig zo'n eenmalig gebruik

uiteindelijk is, valt zeer te betwijfelen. Als woordenboeken nog gemaakt en gebruikt worden, dan zijn het voornamelijk online woordenboeken. Dat geldt ook voor de woordenboeken die gemaakt en ontsloten worden door het Instituut voor de Nederlandse Taal. Deze zijn allemaal online te vinden.[3] Woordenboeken zoals we ze nu nog kennen zullen wellicht te zijner tijd verdwijnen, maar de behoefte aan betrouwbare kennis over woorden en hun betekenissen niet. Die wordt alleen maar groter in de informatiemaatschappij en de big datawereld waarin we nu leven.

Van woordenboek naar computationeel lexicon

We zien steeds vaker dat woordenboeken niet alleen meer gemaakt worden met de menselijke gebruiker in het achterhoofd, maar ook om gebruikt te kunnen worden in taaltechnologische toepassingen. En dat is niet hetzelfde. Computers hebben andere informatie nodig dan mensen. Precies om die reden was het gebruik van machineleesbare woordenboeken voor taaltechnologische toepassingen aan het begin van het digitale tijdperk dan ook niet succesvol. Een klassiek voorbeeld om dit te illustreren, zijn benamingen voor dranken in woordenboeken. Woorden zoals *bier*, *wijn* en *whisky* kunnen verwijzen naar de drank als stofnaam, maar kunnen ook gebruikt worden om te verwijzen naar een portie van de drank. Zo geeft de *Dikke Van Dale Online* (Den Boon & Hendrickx, 2022) voor *whisky* twee betekenissen (stofnaam en glas whisky):


whisky

whis·ky
/ʍɪski/
zelfstandig naamwoord • **de** m • **whisky's** (als stofnaam g.mv.; in Ierland en de VS 'whiskey' gespeld)
1824, Engels ◃ Iers, Gaelisch *uisceabeathadh* [levenswater]

1 sterkedrank, in Schotland uit gerst, in Canada uit mais en rogge gestookt
vergelijk **whiskey**

2 **whisky'tje**
glas whisky (1)

Figuur 1. Woordenboekartikel voor *whisky* (*Dikke Van Dale Online*).

Bij andere dranken, zoals *grappa*, *champagne* en *cider,* wordt deze tweede betekenis 'glas met de drank' niet gegeven in de *Dikke Van Dale Online*. Natuurlijk is het mogelijk om *een grappa, twee champagnes, drie ciders* te zeggen, maar de lexicograaf heeft de extra betekenis bij deze woorden niet opgenomen omdat deze woorden maar zelden of niet in deze betekenis worden gebruikt. Lexicografen beschrijven woorden zoals ze normaal voorkomen, en niet zoals ze mogelijkerwijs ook zouden kunnen voorkomen. Voor mensen hoeft ook niet alles tot in detail gespecificeerd te worden. Mensen begrijpen wel, naar analogie met *whisky*, wat er bedoeld wordt in een zin als *We gaan net aan tafel, ik heb al twee champagnes gedronken*[4]. Voor de computer is dat lastiger. Die wil graag vastgelegd hebben dat dit een systematische variatie is die van toepassing is op alle leden van de set.

Van een lexicon met woorden naar een lexicon met combinaties van woorden

Tot slot dringt niet alleen binnen de lexicografie, maar ook daarbuiten, steeds meer het besef door dat de focus moet verschuiven van de beschrijving van woorden naar de beschrijving van combinaties van woorden. Een mooi voorbeeld daarvan is het project *Woordcombinaties*[5], een online naslagwerk dat geavanceerde leerders en gebruikers van het Nederlands ondersteunt bij het gebruiken van woorden in context.

Neem bijvoorbeeld de woorden *rond* en *tafel* (zie ook Stubbs, 2002:3; Ježek, 2016:24). Het woord *rond*[6] kan 'cirkelvormig' betekenen en *tafel* kan 'meubelstuk, hoofdzakelijk bestaande uit een horizontaal blad, dat op een of meer poten rust, om daarop wat te zetten, te leggen of daaraan wat te verrichten' betekenen. Op basis daarvan kunnen we de combinatie *ronde tafel* vormen voor een tafel die rond is. Echter de combinatie *ronde tafel / rondetafel* heeft ook een volledig andere betekenis, namelijk 'waaraan alle deelnemers als gelijken aanzitten'. Vergelijk *rondetafelconferentie*, *rondetafelbijeenkomst*. Ook kennen we allemaal de ridders van koning Arthur als *de ridders van de ronde tafel*. Onze taalkennis bestaat dus niet alleen uit kennis

van losse woorden, maar vooral uit kennis van combinaties van woorden en de wereldkennis die daarmee samenhangt.

In andere combinaties, betekenen de woorden *rond* en *tafel* namelijk weer iets heel anders (*een rond getal*, *de tafel van vijf*). Dit is voor veel alledaagse woorden het geval. Zonder context kunnen ze verschillende dingen betekenen, maar in context wordt de betekenis meestal al snel duidelijk. Het idee dat niet de afzonderlijke woorden, maar juist 'woorden in context' centraal moeten staan, is zeker niet nieuw en is al terug te vinden in werk van o.a. Harris, Firth, Mel'čuk en Sinclair.

John Sinclair, een van de grote pioniers van de corpuslinguistiek en -lexicografie, zag een tegenstelling tussen wat hij het 'idiom principle' en het 'open choice principle' noemt (1991:110-115). Het 'idiom principle' stelt dat taalgebruikers een groot aantal semivoorgeconstrueerde zinnen of zinsdelen tot hun beschikking hebben, die als een geheel functioneren, terwijl het 'open choice principle' stelt dat de meeste woorden in een taal gecombineerd kunnen worden met de meeste andere woorden in de taal. Een werkwoord zoals *accepteren*, kan met veel woorden gecombineerd worden (bijvoorbeeld *situatie*, *gift*, *voorstel*, *erfenis*, etc.). Een werkwoord als *stipuleren* daarentegen komt slechts in een beperkt gezelschap van andere woorden voor en wordt vooral met *voorwaarde* gebruikt.

Volgens sommige schattingen zou ongeveer 40% tot 60% van een tekst uit dergelijke prefabs bestaan (Erman & Warren, 2000). Volgens Mel'čuk (2023:3) is dit percentage nog hoger en komen wat hij 'holistic multilexemic expressions' noemt, 10 keer zo vaak voor als losse woorden. Mel'čuk (1995:169) stelt dan ook: 'People do not speak in words; they speak in phrasemes'. Wat de omvang van dit taalkundige fenomeen precies is, hangt natuurlijk af van de gehanteerde definitie van zo'n prefab, en alhoewel dit zeer interessant is, valt het buiten het bestek van deze oratie.

Het gaat om een taalkundig verschijnsel dat wijdverspreid is. Het is dan ook niet verwonderlijk dat de studie van woordcombinaties de aandacht heeft getrokken van onderzoekers uit verschillende vakgebieden (o.a. computerlinguïstiek, lexicografie, theoretische en toegepaste taalkunde). Ieder met een andere invalshoek en tot voor kort met weinig overlap ertussen. Binnen de computerlinguïstiek houdt men zich vooral bezig met de automatische verwerking, m.n. het opsporen en identificeren van combinaties van woorden in teksten. Als de computer niet weet dat iets een vaste combinatie is, dan kan hij het bijvoorbeeld ook niet correct vertalen. Binnen de theoretische taalkunde is er vooral aandacht voor de specifieke eigenschappen van woordcombinaties zoals de mate van vastheid en idiomaticiteit, terwijl onderzoekers binnen de toegepaste taalkunde juist weer geïnteresseerd zijn in de rol die fraseologie speelt bij taalverwerving en didactiek.

Fraseologie

De verschillende werelden komen nu steeds dichter bij elkaar en het vakgebied waarin ze samenkomen, de (computationele) fraseologie, is booming (zie Corpas Pastor & Colson, 2020). Daar waar de fraseologie zich eerder vooral bezighield met de vaste en ondoorzichtige combinaties van woorden (uitdrukkingen zoals *het vat der Danaïden vullen*[7]), bestrijkt het vakgebied nu een veel breder scala aan combinaties.

Het moment lijkt nu dus rijp voor een kruisbestuiving tussen de verschillende benaderingen, maar de gezamenlijke studie wordt nog enigszins gecompliceerd door het ontbreken van eenduidige terminologie. Zoals Mel'čuk (2023:5) stelt: '[…] linguistics is not known for its strict terminology, and phraseology is probably the worst terminology-served domain'. We zien dat de computerlinguïstiek een duidelijke voorkeur heeft voor de term 'multiword expression' (MWE), terwijl in de taalkunde vaker over fraseologische eenheid, fraseem of vaste verbinding wordt gesproken. Maar ook voor deze termen hanteert niet iedere onderzoeksgroep dezelfde definitie wat de afbakening van het domein lastig maakt. Eenduidige terminologie is erg belangrijk (zie onder andere Tiberius et al., 2021 over lexicografische terminologie). Ik wil me vandaag echter niet te veel

bezighouden met terminologie (want dan zijn 45 minuten zeker niet genoeg), en ik zal daarom de volgende praktische omschrijving hanteren om het scala aan lexicale items dat binnen de fraseologie bestudeerd wordt te positioneren:

> 'Combinaties van (orthografische) woorden die om een of andere reden als eenheid moeten worden opgeslagen in computationele lexica of woordenboeken.' (gebaseerd op Al Haj et al., 2013:130)[8]

Het woord 'orthografisch' staat hier met opzet tussen haakjes. De meeste mensen denken namelijk bij het begrip 'woord' intuïtief aan orthografische woorden, begrensd door spaties. Echter, als we uitgaan van de spelling, dan zouden samenstellingen zoals *computerlinguïstiek* en *hoogleraar* buiten de boot vallen omdat ze aaneengeschreven worden. Daarom wordt voor talen als het Nederlands (met veel samenstellingen) spelling meestal niet als een strikt criterium gebruikt om het domein van de fraseologie af te bakenen.

Om het allemaal wat concreter te maken zal ik nu proberen de verscheidenheid aan combinaties te illustreren aan de hand van een stukje tekst van een puzzelwandeling in Leiden[9].

Wetenschap deelt de lakens uit
Leiden | 4 km
Start- en eindpunt: NS Leiden Centraal
Alstublieft, uw wandelroute met puzzels!
Leiden, wat een heerlijke stad! Deze stad ademt kwaliteit, deze stad ademt onderzoek en wijsheid. Door de eeuwen heen heeft Leiden een vooraanstaande rol kunnen spelen op meerdere terreinen. Waar de talrijke musea ons meenemen naar die periode, kijken universiteit en onderzoekers alweer vooruit. Het maakt Leiden dynamisch en bruisend. Wij wandelen door de geschiedenis en combineren dat met het heden. Je kijkt je ogen uit! Op de route 5 puzzels. Kijk op pagina 4 en los ze allemaal op.
Veel wandel- en puzzelplezier!

Deze tekst bevat verschillende typen combinaties. Sommige zijn semantisch transparant, zoals *dynamisch* en *bruisend* in combinatie met *stad*, bij andere kan de betekenis niet zonder meer uit de samenstellende delen worden afgeleid zoals in *de lakens uitdelen* en *je ogen uitkijken*. Sommige combinaties zijn (volledig) vast, de woorden moeten in een vaste volgorde staan en er is geen modificatie mogelijk. Een voorbeeld van zo'n vaste combinatie is *door de eeuwen heen*. Eigennamen zijn ook vast zoals *NS Leiden Centraal*. De meeste combinaties zijn echter niet volledig vast. In de combinatie *pagina 4* is de volgorde wel vast, maar naast *4* kan hier elk willekeurig getal voorkomen. In de tekst is ook de formule *veel plezier!* uitgebreid tot *veel wandel- en puzzelplezier!* Er kunnen dus andere woorden tussen de vaste delen van sommige combinaties staan. Dat zien we ook bij de scheidbaar samengestelde werkwoorden *vooruitkijken* en *oplossen* in deze tekst. Een mooi voorbeeld van een zinspatroon is te vinden in de zin *deze stad ademt kwaliteit*, waar het werkwoord *ademen* metaforisch wordt gebruikt. Tot slot zien we dat combinaties zelf ook weer gecombineerd kunnen worden. Zo kunnen de combinaties *vooraanstaande rol* en *rol spelen* gecombineerd worden tot *vooraanstaande rol spelen*.

Wat al deze woordcombinaties met elkaar verbindt is de coselectie van woordenschat en grammaticaregels om een bepaalde betekenis in een bepaalde sociale situatie uit te drukken (Stubbs, 2009:120). Hierdoor staan ze succesvolle automatische verwerking in de weg die op deze traditionele tweedeling is gebaseerd (zie Vondřička, 2019; Sag et al., 2002). Is de combinatie volledig versteend en staat deze in het lexicon, dan is er in principe geen probleem. Echter, vaak komt de vorm van een combinatie in een tekst niet precies overeen met de vorm van de combinatie zoals deze in het woordenboek of lexicon is opgenomen. Veel combinaties vertonen een zekere flexibiliteit. Zo kan men in het Engels zowel *grasping at straws* zeggen als *clutching at straws* 'zich aan iedere kleinigheid vastklampen'. *Straws* hoeft niet noodzakelijk in het meervoud te staan. Het is ook mogelijk om *to clutch/grasp at a straw* te zeggen en het werkwoord kan ook na het zelfstandig naamwoord voorkomen

in *the result seemed but a straw to clutch at*.[10] Al deze variatie maakt automatische identificatie en extractie van dergelijke combinaties moeilijk. Daarom wordt er in de computerlinguïstiek steeds vaker uitgegaan van een canonieke vorm van de combinatie (vergelijkbaar met het trefwoord in een woordenboek) en worden speciale codes gebruikt om restricties en variatie aan te geven. In combinatie met de syntaxis wordt het dan mogelijk om de verschillende varianten van een combinatie automatisch te identificeren. Een mooi voorbeeld hiervan is de MWE-Finder[11] (Odijk et al., te verschijnen).

Verder zijn veel combinaties niet compositioneel van aard en kan de betekenis niet uit de samenstellende delen worden afgeleid. Computationele methoden zijn daarentegen juist vaak wel compositioneel en kunnen hier dus niet zo goed mee omgaan (Savary et al., 2019). Zo herkende het automatische vertaalsysteem van de Europese Commissie, eTranslation[12], de uitdrukking *de lakens uitdelen* niet in de titel van de tekst bij de Leidse puzzelwandeling en vertaalde de titel in het Engels als *Science distributes the sheets*.

Een ander aspect dat de computationele verwerking in de weg staat, en dat hiermee samenhangt, is de lage frequentie van veel combinaties (vooral de meer idiomatische). Zo komt de combinatie *lakens uitdelen* slechts 3 keer voor in een verzameling van bijna 9 miljoen woorden (gesproken taal)[13], en zo'n 3800 keer in een verzameling van bijna 3 miljard woorden (geschreven taal)[14]. Dit in tegenstelling tot de combinatie *rol spelen* die net iets meer dan 165.000 keer voorkomt in de grootste verzameling.

Voor hoogfrequente combinaties - combinaties van woorden die bovengemiddeld vaak samen voorkomen zoals *rol spelen* - bestaan verschillende statistische associatiematen en machinelearningtechnieken waarmee voor de automatische verwerking ervan goede resultaten kunnen worden bereikt. Maar die werken niet voor laagfrequente combinaties omdat ze simpelweg niet vaak genoeg voorkomen. De combinatie *lakens uitdelen* gedraagt zich syntactisch net als een regelmatige combinatie, *kaartjes uitdelen*, maar komt in een corpus maar zeer zelden voor. Voor een computer is het dan moeilijk om het verschil te leren tussen deze twee. Kortom, woordcombinaties vormen nog steeds een 'Pain in the Neck for NLP' (Sag et al., 2002). Om de fraseologie in de computerlinguïstiek een boost te geven zijn met name twee componenten belangrijk: corpora en computationele lexica.

Corpora

Een corpus is een verzameling taaldata in elektronische vorm. Het kan bestaan uit geschreven materiaal, gesproken materiaal, gebarentaal, het kan teksten uit kranten of boeken bevatten uit verschillende perioden en evt. ook in verschillende talen. Zo bieden corpora onderzoekers de mogelijkheid om op systematische wijze te kijken naar grote hoeveelheden authentiek taalgebruik, geproduceerd door 'echte' mensen. Daarmee vormen ze een nuttige en belangrijke bron van bewijsmateriaal voor linguïstisch en computerlinguïstisch onderzoek. Zoals Kilgarriff stelt:

> "corpus data allow us to study language 'with a degree of objectivity ...where before we could only speculate'" (Kilgarriff, 1997:137).

De laatste paar decennia is de beschikbaarheid van digitale data enorm toegenomen. Als gevolg daarvan worden corpora steeds groter. Voor het Nederlands is er het Corpus Hedendaags Nederlands (CHN), dat op dit moment al meer dan 9.250.000 teksten bevat uit kranten, tijdschriften, journaaluitzendingen, blogs en boeken uit Nederland en de Caribische rijksdelen, België en Suriname. Een corpus van dergelijke omvang is goed nieuws voor de studie van woordcombinaties. We zagen al dat grote hoeveelheden data nodig zijn om voldoende informatie te krijgen over de laagfrequente combinaties. Het is echter belangrijk om u te realiseren dat hoe groot een corpus ook is, het blijft altijd slechts een steekproef van daadwerkelijk taalgebruik. Het feit dat een bepaalde uitdrukking of construc-

tie niet in het corpus voorkomt, betekent dan ook niet automatisch dat de uitdrukking of constructie niet bestaat. Veel verschijnselen in natuurlijke taal zijn laagfrequent of zijn laagfrequent in een bepaalde variëteit (de wet van Zipf). Formules, zoals *goedemorgen,* komen vooral in gesproken taal voor en treffen we meestal niet veelvuldig aan in geschreven teksten.

Omgekeerd is ook niet alles wat in een corpus staat correct. Corpora bevatten authentieke teksten, geschreven door mensen. Mensen maken fouten en die fouten zitten dus ook in het corpusmateriaal. Zo wordt in het Corpus Hedendaags Nederlands[15] meerdere malen *het ei van Columbus uitgevonden*, terwijl dat eigenlijk *het warm water uitvinden* of *het wiel uitvinden* moet zijn.

In het ideale geval bevat een corpus een gebalanceerde verzameling van taaldata die representatief zijn voor de taal of taalvariëteit die men wil onderzoeken. Het kan echter voorkomen dat er te veel materiaal van een bepaald type in het corpus zit. In dat geval spreken we van bias. Bias zien we veel in webcorpora. Niet alle sprekers van een taal produceren evenveel materiaal op het internet. Bepaalde groepen zijn oververtegenwoordigd, andere juist ondervertegenwoordigd of zelfs niet vertegenwoordigd. Volgens onderzoek naar het geslacht van mensen die bijdragen leveren aan Wikipedia, is slechts 8-15% van de Wikipedianen, vrouw (zie ook Barera, 2020).

Resultaten moeten daarom altijd geïnterpreteerd worden in de context van het corpus dat gebruikt is. Goede metadata, d.w.z. informatie over de teksten die in het corpus zitten, zijn hiervoor essentieel. Wie is de auteur, wanneer is de tekst geschreven, wat is het domein, etc. Hoe gedetailleerder de metadata, hoe specifieker de resultaten geïnterpreteerd kunnen worden.

Naast het toevoegen van metadata, wordt een corpus normaal gesproken ook taalkundig verrijkt. Dat betekent dat aan de woorden in het corpus taalkundige informatie wordt toegevoegd, zoals woordsoort, lemmavorm en een syntactische structuur.

```
# text = Wij wandelen door de geschiedenis.
1   Wij          wij          PRON   2   nsubj
2   wandelen     wandelen     VERB   0   root
3   door         door         ADP    5   case
4   de           de           DET    5   det
5   geschiedenis geschiedenis NOUN   2   obl
6   .            .            PUNCT  2   punct
```

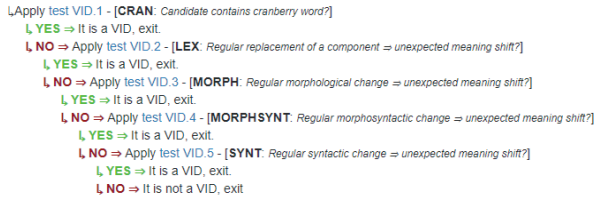
Figuur 2. Voorbeeld van taalkundige verrijking met UDPipe[16] in CoNLL-U-formaat[17].

Met behulp van deze extra informatie, kunnen verschillende onderzoeksvragen beter en sneller onderzocht worden. Het taalkundig verrijken van corpora gebeurt tegenwoordig meestal automatisch, maar om de automatische analyse verder te kunnen verbeteren zijn handmatig geannoteerde corpora ook nog steeds nodig. In dit kader wil ik graag een actueel initiatief noemen.

Een PARSEME-corpus voor het Nederlands

PARSEME[18] is een internationaal initiatief waarbinnen corpora worden ontwikkeld waarin verschillende typen combinaties met een werkwoord (Verbal Multiword Expressions) geannoteerd zijn, zoals *prendre la poudre d'escampette* 'het hazenpad kiezen' (VID), *eine Entscheidung treffen* 'een beslissing nemen' (LVC), *laisser tomber* 'opgeven' (MVC), *zich vergissen* (IRV) en *to take off* 'opstijgen' (VPC). Op dit moment zijn er PARSEME-corpora voor 26 verschillende talen, maar het Nederlands ontbreekt nog. In een samenwerkingsverband tussen de universiteiten van Groningen, Leiden en Utrecht, en het Instituut voor de Nederlandse Taal hopen we in 2024 zo'n corpus voor het Nederlands te realiseren (Bouma et al., 2024).
Voor het annoteren van de combinaties zijn door PARSEME universele richtlijnen[19] opgesteld, die gedefinieerd zijn in de vorm van een beslisboom met ja-neevragen en specifieke tests. Het doorlopen van deze beslisboom vergt een gedegen taalkundige kennis van het Nederlands. Bijvoorbeeld om vast te stellen of een combinatie een zogeheten 'verbal idiom' (VID)

is, moet eerst gekeken worden of de combinatie een 'cranberry word' bevat, dat wil zeggen een woord dat niet los voorkomt zoals het Franse woord *martel* (*se mettre martel en tête* 'zich veel zorgen maken'). Als dat niet het geval is, moet achtereenvolgens gekeken worden of lexicale, morfologische, morfosyntactische of syntactische variatie mogelijk is zonder dat de betekenis van de combinatie verandert. Zodra het antwoord op een van deze vragen positief is, kunnen we de boom verlaten.



Figuur 3. Deel van de PARSEME-beslisboom voor 'verbal idioms'.

Het mooie aan het PARSEME-initiatief is dat de corpora niet alleen ingezet kunnen worden voor het trainen, tunen en testen van nieuwe algoritmen voor het automatisch detecteren van woordcombinaties, maar dat ze ook crosslinguïstisch onderzoek faciliteren en zo bijdragen aan meer inzicht in de diversiteit en de universaliteit van dit specifieke type combinaties.

Computationele lexica

Een tweede belangrijke component voor succesvolle computationele verwerking van fraseologie is een lexicon. Lexica bevatten informatie over woorden en hoe ze worden gebruikt en zijn daarmee complementair aan corpora. Veel bestaande lexica voor woordcombinaties zijn echter onvoldoende geformaliseerd waardoor ze niet zonder meer in te zetten zijn voor computerlinguïstische toepassingen (Savary et al., 2019). Daarnaast ligt in de meeste computationele lexica de nadruk op de morfosyntactische eigenschappen van de combinaties. Codering van de semantische eigenschappen vormt nog een belangrijke uitdaging. Ik wil daar graag onderzoek naar doen. Een interessant uitgangspunt voor dit onderzoek, is te vinden in lexicografische hoek, in het werk van Patrick Hanks (een leerling van John Sinclair en net als Sinclair geïnspireerd door de distributionele semantiek).

Theory of Norms and Exploitations

Volgens Hanks (2013) hebben (losse) woorden geen betekenis maar slechts betekenispotentieel. Betekenis wordt geactiveerd als woorden gecombineerd worden met andere woorden, in zinnen. In de zin *De bakker bakt brood* wordt een heel andere betekenis van het werkwoord *bakken* geactiveerd dan in de zin *De bakker bakt er niets van*. Deze semantisch gemotiveerde terugkerende structuren van woorden, noemt Hanks patronen. In zijn Theory of Norms and Exploitations (TNE)[20] onderscheidt hij twee soorten patronen: de normale, prototypische patronen van woorden, de normen, en creatief gebruik van die normale patronen, de exploitaties. Exploitaties ontstaan wanneer onverwachte woorden of onverwachte metaforen worden gebruikt. Illustratief voor een normaal patroon voor het Engelse werkwoord *to talk* is deze zin waarin iemand spreekt met iemand anders over iets:

> 'I want to talk to you about that.'[21]

Een mooi voorbeeld van een exploitatie van het werkwoord *to talk* is te vinden in een uitspraak van de golfer Lee Trevino:

> 'As Lee Trevino says: "You can talk to a fade but a hook just won't listen."'[22] (zie ook Hanks 2013:213)

Deze zin is alleen voor de golfers onder ons te begrijpen (en met name diegenen die af en toe tegen hun golfbal praten). De woorden *fade* en *hook* zijn niet van het type [[Human]] dat normaal gesproken bij *to talk* gebruikt wordt, maar specifieke golftermen. Bij een *fade* wordt de bal geslagen met een bepaalde balvlucht, een *hook* is een complete afzwaaier.

Normen en exploitaties zijn twee uitersten van het spectrum en er is geen scherpe scheidingslijn. Sommige normen zijn normaler dan andere; sommige exploitaties zijn extremer dan andere. Tussen de normen en de exploitaties in vinden we varianten. Varianten kunnen onder andere lexicaal, syntactisch of semantisch van aard zijn. We zagen eerder al dat in plaats van het werkwoord *to grasp* in *grasping at straws* het werkwoord *to clutch* gebruikt kan worden zonder dat dit de betekenis substantieel verandert. Dit is een voorbeeld van lexicale variatie. Syntactische variatie, zoals actief/passief, komt ook vaak voor. Veel werkwoorden kunnen zowel in het actief als in het passief voorkomen zonder dat dit invloed heeft op de basisbetekenis. Tot slot, kan semantische variatie zich voordoen. Zo kunnen werkwoorden waarbij het onderwerp van het type [[Human]] is, meestal ook gebruikt worden met het type [[Human Institution]]. Bijvoorbeeld, *onderzoekers kijken alweer vooruit* versus *de universiteit kijkt alweer vooruit*. Dit is een regelmatige alternantie en vormt geen afzonderlijk patroon.

Om inzicht te krijgen in de normen en exploitaties van een taal bestudeert TNE daadwerkelijk taalgedrag zoals dat is vastgelegd in corpora. De theorie biedt praktische richtlijnen in de vorm van Corpus Pattern Analysis (CPA) om deze taaldata te sorteren en classificeren. Collocaties spelen in deze analyse een belangrijke rol. Nemen we bijvoorbeeld het werkwoord *boeken*. We kunnen *inkomsten, omzet, verlies* en *winst* boeken en dat betekent dan 'iets in de boekhouding schrijven'. Daarnaast kan er ook *vooruitgang, resultaat, overwinning* en *succes* geboekt worden. In combinatie met deze zelfstandige naamwoorden wordt de betekenis 'iets behalen' geactiveerd. Woorden als *overnachting, hotel, appartement*, etc. activeren nog een andere betekenis, namelijk 'iets reserveren'. We zien drie clusters van zelfstandige naamwoorden die steeds een andere betekenis activeren. Zo'n cluster noemen we een lexicale set. Aan die lexicale set kan dan weer een semantisch type gekoppeld worden, namelijk [[Money Value]] voor het eerste cluster met *inkomsten*, [[Goal]] voor het tweede cluster met *vooruitgang*, en [[Event, Location]] voor het cluster met *overnachting* en *hotel*. De semantische typen zijn vastgelegd in een hiërarchisch gestructureerde ontologie. De ontologie van Hanks bevat zo'n 250 verschillende semantische typen (Ježek & Hanks, 2010).[23] Voor de leesbaarheid kunnen de semantische typen vervangen worden door dummies zoals *iemand* of *iets* in de resulterende patronen.

1	**iemand**	**boekt**	**iets**
	[[Human, Institution]]		**[[Money Value]]**
	bank, bedrijf, beurs, …		*inkomsten, omzet, winst,…*
	Iemand schrijft iets bij in de boekhouding		
	Het bedrijf boekte een omzet van 53,78 miljard euro.		

2	**iemand**	**boekt**	**iets**
	[[Human, Institution]]		**[[Goal]]**
	iemand		*overwinning, resultaat, succes, vooruitgang, …*
	Iemand behaalt iets		
	De jongste jaren heb ik veel vooruitgang geboekt.		

3	**iemand**	**boekt**	**iets** of **iemand**
	[[Human, Institution]]		**[[Event, Location, Document, Human, Human Group]]**
	iemand		*appartement, artiest, hotel, kamer, overnachting, reis, ticket, …*
	Iemand reserveert iets of iemand		
	Hoe eerder je een ticket boekt, hoe goedkoper dat is.		

Figuur 4. Patronen voor het werkwoord *boeken* (zie ook *Woordcombinaties*).

Gebruikelijke uitdrukkingen worden als afzonderlijke patronen opgenomen.

4	**Uitdrukking**		
	iemand of **iets**	**staat geboekt**	als **iemand** of **iets**
	[[Entity]]		
	iemand of iets staat bekend als iemand of iets		
	De N-VA-voorzitter staat niet geboekt als een flapuit.		

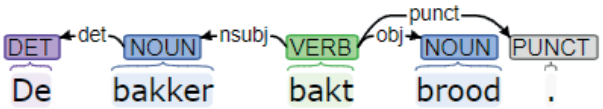
Figuur 5. Patroon van de uitdrukking *geboekt staan als* (zie ook *Woordcombinaties*).

Op het eerste gezicht lijkt patroonanalyse misschien eenvoudig, maar niets is minder waar. De praktijk leert dat hoe meer voorbeelden bekeken worden van een woord, hoe onduidelijker het meestal wordt. Dezelfde syntactische structuur kan in meerdere patronen voorkomen omdat verschillende betekenissen worden geactiveerd (zoals bij *boeken* het geval is). Verschillende syntactische structuren kunnen echter ook dezelfde betekenis hebben. In zo'n geval moet worden vastgesteld of het om een syntactische variant gaat of om een ander patroon. Net als het annoteren van combinaties binnen PARSEME, vereist CPA gedegen taalkundige en lexicografische expertise.

TNE wordt in de praktijk getoetst in het project *Pattern Dictionary for English Verbs*[24]. Ook voor het Spaans (Renau & Nazar, 2016), Italiaans (Ježek et al., 2014), Kroatisch (Marini & Ježek, 2019) en het Nederlands (Colman & Tiberius, 2018) zijn er vergelijkbare initiatieven. Voor het Nederlands worden patronen bewerkt in de context van het project *Woordcombinaties*.

Het maken van dergelijke patroonwoordenboeken is nu nog hoofdzakelijk een computerondersteund handmatig proces en bijgevolg bijzonder tijdrovend. De collocaties worden semiautomatisch geëxtraheerd, maar het annoteren van de patronen gebeurt in de meeste projecten volledig handmatig. Ik wil een aantal strategieën voor automatisering van het proces voorstellen.

Patronen zijn syntactisch-semantische structuren. Het corpus dat gebruikt wordt voor de analyse, is idealiter dan ook verrijkt met zowel een syntactische als een semantische annotatielaag. Voor de syntactische analyse kan uitgegaan worden van Universal Dependencies (UD; de Marneffe et al., 2021)[25] een universeel 'framework' voor consistente annotatie van grammatica (woordsoortinformatie, morfologische kenmerken en syntactische dependenties). Figuur 6 geeft een UD-annotatie van de zin *De bakker bakt brood*. Voor het parseren met UD bestaan verschillende tools[26], die relatief eenvoudig in de workflow te integreren zijn. De foutjes die hier en daar nog in de analyse sluipen nemen we daarbij voor lief.



Figuur 6. UD-annotatie[27].

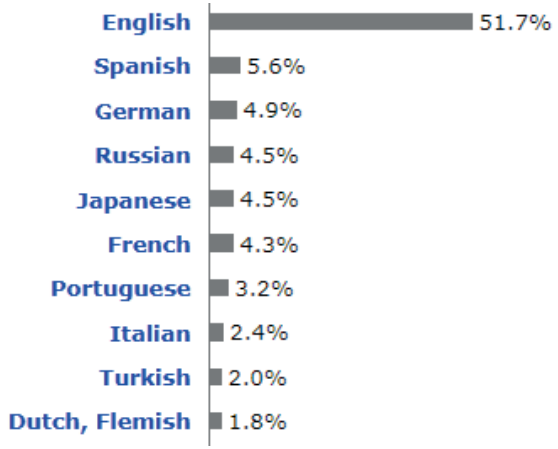
Naast een syntactische analyse is ook een semantische analyse nodig in de vorm van semantische typen die binnen CPA worden onderscheiden (d.w.z. *bakker* is van het type [[Human]] en *brood* is van het type [[Food]]). Om dit automatisch te kunnen doen, moet de computer geleerd hebben dat woorden zoals *brood*, *taart*, *pannenkoek*, etc. van het semantische type [[Food]] zijn. Dit is niet zo eenvoudig omdat veel woorden tot meerdere semantische typen kunnen behoren (bijvoorbeeld het woord *vis* kan zowel van het type [[Animal]] als [[Food]] zijn, maar dat geldt dan weer niet voor het woord *bokking* dat alleen van het type [[Food]] is). Voor het Nederlands bestaan er nog geen kant-en-klare oplossingen voor het semantisch labelen met de CPA-ontologie. De USAS[28] parser van Lancaster komt in de buurt en is dankzij een toegankelijke Python-implementatie[29] eenvoudig te gebruiken. Het Nederlandse lexicon van de parser is echter beperkt tot de 5000 meest frequente woorden van het Nederlands (Tiberius & Schoonheim, 2014). Logischerwijze worden een heleboel woorden dan ook niet herkend tijdens het parseren. Daarnaast komt de indeling in semantische klassen die deze parser gebruikt niet overeen met de semantische typen in de CPA-ontologie en moet er eerst gekoppeld worden. Een andere mogelijkheid is om te vertrekken vanuit de CPA-ontologie en deze semiautomatisch te vullen met woorden uit het corpus met behulp van verschillende distributionele technieken. Dit is onderzocht in het kader van het project *Pattern Dictionary for Spanish Verbs* (Verbario)[30] (Nazar & Renau, 2016; Renau & Nazar, 2016). Met de resulterende ontologie worden woorden in het Spaanse corpus automatisch semantisch getagd en op basis daarvan worden automatisch patronen geëxtraheerd (met een precisie tussen de 40% en 50%[31]). Eenzelfde techniek zouden we ook voor het Nederlands kunnen toepassen. Echter, met de hoeveelheid handmatige gelabelde data die reeds binnen de verschillende patroonwoordenboeken is gecreëerd, is machinelearning en in het bijzonder 'supervised machine learning' tegenwoordig een realistisch alternatief voor het ontwikkelen van een dergelijke semantische parser. Op basis van de handmatig gelabelde data kunnen we de computer trainen en hem leren om semantische typen toe te kennen aan de woorden in het corpus. Ook kunnen we de computer trainen op voorbeeldzinnen van patronen die handmatig zijn geannoteerd en vervolgens van hem verwachten dat hij zinnen die nog niet geclassificeerd zijn onder het juiste patroon weet te hangen. Onderzoek zal moeten uitwijzen hoe accuraat we computers op deze taken kunnen trainen.

Tot slot, mogen gezien de huidige ontwikkelingen, de LLM's, grote taalmodellen, natuurlijk niet in het rijtje van mogelijke automatiseringsstrategieën ontbreken. LLM's zijn gigantische kunstmatige neurale netwerken die getraind zijn op basis van miljarden stukjes data. Op basis hiervan leren ze zelf verbanden te leggen en zo kunnen ze onder andere taal reproduceren. Er bestaat een groot aantal verschillende taalmodellen, die door middel van wat men 'transfer learning' noemt, gefinetuned kunnen worden op een specifieke taak (bijvoorbeeld het beantwoorden van vragen zoals in het geval van ChatGPT). Om een eerste indruk te krijgen van wat deze modellen voor CPA kunnen betekenen, hebben we gekeken naar ChatGPT[32]. We hebben de chatbot gevraagd om teksten syntactisch en semantisch te annoteren en patronen voor werkwoorden te genereren. De voorlopige resultaten suggereren dat syntactisch annoteren relatief goed gaat, maar dat de andere twee taken nog erg lastig zijn. Zo bakt ChatGPT er nog niet zoveel van als we de chatbot vragen om patronen voor het werkwoord *bakken* te genereren. ChatGPT bakt van alles, van cake en koekjes tot pannenkoeken en flensjes, met liefde en zonder recept allemaal in afzonderlijke patronen die maar weinig overeenkomst vertonen met de patronen die handmatig voor *bakken* zijn vastgesteld. Meer onderzoek is nodig. Kan betere prompting tot betere resultaten leiden, is training en finetuning noodzakelijk of moeten we een taalmodel gebruiken dat specifiek getraind

is op het Nederlands? Bijkomend probleem is dat we met de huidige taalmodellen de link met het onderliggende corpus kwijt zijn en daardoor de resultaten niet eenvoudig te valideren zijn. Desalniettemin liggen hier heel wat interessante en nog onbeantwoorde onderzoeksvragen.

Besluit

Ik kom tot een besluit. Met behulp van corpora, corpustools en computationele technieken is het nu mogelijk om op grote schaal te onderzoeken hoe woorden worden gebruikt, hoe ze samengaan en hoe ze zo betekenis 'maken'. Een dergelijke systematische analyse van grote hoeveelheden corpusdata in verschillende talen zal leiden tot betere lexica. Deze zullen op hun beurt weer leiden tot betere tools voor natuurlijke taalverwerking door de computer. Sommige aspecten van bestaande taalkundige theorieën zullen worden bevestigd door deze analyse, andere zullen juist worden weerlegd. Als zodanig verenigt Computational Corpus Pattern Analysis beide toepassingen van de computerlinguïstiek: het toetsen van een theorie, in het bijzonder TNE, en het realiseren van een groot computationeel lexicon van prototypische fraseologische patronen en hun betekenissen. De inzet van taalmodellen biedt hierbij nieuwe mogelijkheden die verder onderzocht moeten worden. Het gebruik van taalmodellen heeft echter ook een keerzijde. Deze modellen worden vooral getraind op materiaal dat afkomstig is van het internet en we zagen al dat webcorpora gevoelig zijn voor bias. Niet alle groepen mensen zijn gelijk vertegenwoordigd op het internet. Dit geldt ook voor de talen op het internet. De meeste teksten op het internet zijn in het Engels. De percentages voor de andere talen zijn een stuk lager, en voor bedreigde talen of dialecten zijn deze nog veel lager. Speciale technieken en strategieën zijn nodig om ervoor te zorgen dat deze low-resourcetalen, waar de hoeveelheid trainingsdata schaars is, niet buitengesloten worden in de digitale wereld.



Figuur 7. Percentages van websites (met inhoud) in verschillende talen (W3Techs.com, 30 januari 2024)

Ook is het belangrijk om te onderzoeken in welke mate de dominantie van het Engels in de trainingsdata van de grote taalmodellen een effect heeft op de resultaten in andere talen. Zo bevatten de resultaten in andere talen soms leenwoorden en leenvertalingen, maar de hoofdzakelijk Engelse input lijkt ook de grammatica in de output te beïnvloeden. Dit kan uiteindelijk leiden tot een nivellering van structurele grammaticale verschillen tussen talen. Dit potentiële nivellerende effect wordt verder versterkt als machinaal geproduceerde teksten weer gebruikt worden als input voor het trainen van de volgende generatie taalmodellen. Barrett (2023, geciteerd in De Schryver, 2023:362) spreekt in dit verband over het gevaar van inteelt en monocultuur.

Dit alles maakt onderzoek naar de prototypische fraseologische patronen in verschillende talen en genres, niet alleen interessant, maar ook noodzakelijk. Zoals lichtvervuiling en overvolle steden, het kijken naar sterren tegenwoordig moeilijk, zo niet onmogelijk maken, zo wordt het bestuderen van patronen in natuurlijke talen lastig als een steeds groter aandeel van de

teksten niet meer door mensen wordt geschreven maar gegenereerd wordt door machines.

Fraseologie is een interdisciplinair vakgebied. Om fraseologie volledig te doorgronden, is het nodig om de inzichten uit de verschillende vakgebieden (o.a. computerlinguïstiek, lexicografie, linguïstiek, psycholinguïstiek, en onderwijswetenschappen) samen te brengen. Al deze vakgebieden zijn in Leiden vertegenwoordigd en ik hoop dan ook dat mijn onderzoek tot veel nieuwe samenwerkingen zal leiden. Op deze manier hoop ik met mijn leerstoel een positieve bijdrage te leveren aan 'linguistic data science' binnen het LUCL.

Dankwoord

Aan het eind van deze oratie gekomen, dank ik het College van Bestuur van de Universiteit Leiden en de leden van de benoemingscommissie van harte voor mijn benoeming en het in mij gestelde vertrouwen. Ik voel mij bevoorrecht dat ik mijn onderzoek aan de Universiteit Leiden mag uitvoeren.

Wetenschappelijk onderzoek doe je niet alleen, maar berust op teamwerk. Er is helaas geen ruimte om iedereen te noemen met wie ik de afgelopen jaren in nationale en internationale projecten heb samengewerkt en die mij hebben geïnspireerd, gesteund en gemotiveerd in welke vorm dan ook. Ik ben jullie daar allemaal zeer erkentelijk voor.

Ik wil mijn collega's van het LUCL bedanken voor de hartelijke wijze waarop ze mij in hun midden hebben opgenomen. Ook wil ik mijn collega's van het INT bedanken voor de prettige werkomgeving en hun collegialiteit, in het bijzonder Lut Colman. Ik hoop dat mijn benoeming zal bijdragen aan een nauwere samenwerking tussen het instituut en de universiteit.

Ik wil graag een aantal mensen die een belangrijke rol gespeeld hebben in mijn academische loopbaan in het bijzonder bedanken.

Hooggeleerde Cheng, beste Lisa en Hooggeleerde Steurs, beste Frieda. Zonder jullie stond ik hier nu niet. Ik wil jullie bedanken voor het vertrouwen dat jullie in mij hebben.

Hooggeleerde Corbett, beste Grev. Ik heb altijd met heel veel plezier deel uitgemaakt van jouw onderzoeksgroep, de Surrey Morphology Group. Het werk en de discussies hebben mijn kennis van de taalkunde enorm verdiept.

Hooggeleerde Brown, beste Dunstan. Ik wil jou in het bijzonder bedanken. Het was een voorrecht om met jou te mogen samenwerken en ik prijs mij gelukkig dat we na al die jaren nog steeds contact hebben. Je onuitputtelijke enthousiasme voor de wetenschap werkt aanstekelijk.

Hooggeleerde Gazdar, beste Gerald. Hooggeleerde Evans, beste Roger. Ik had me geen betere promotoren kunnen wensen. Ik heb enorm veel van jullie geleerd en ik denk nog altijd met veel plezier terug aan mijn promotietijd in Brighton.

Hooggeleerde Coppen, beste Peter-Arno. In Nijmegen liggen mijn computationele roots met de opleiding Taal, Spraak en Informatica. Dank voor het leggen van een goede basis.

Hooggeleerde D'Hulst, beste Lieven. In Antwerpen is het allemaal begonnen. Ik denk nog steeds dankbaar terug aan mijn licentiaatsonderzoek over collocaties dat ik onder jouw begeleiding mocht uitvoeren.

Een aparte vermelding verdienen ook de studenten. Ik geef jullie met veel plezier college en geniet iedere keer weer van jullie energie. Ik verheug me dan ook op het verder vormgeven van de track Computational Linguistics op bachelor- en masterniveau.

Voordat ik afsluit, wil ik even stilstaan bij een aantal collega's die dit moment helaas niet meer mogen meemaken: mijn derde promotor en grote inspirator, Adam Kilgarriff, mijn

collega en co-auteur van het frequentiewoordenboek, Tanneke Schoonheim, en Patrick Hanks, die onlangs is overleden. Zij worden gemist.

Familie en vrienden, ik ben zeer dankbaar dat jullie hier vandaag aanwezig zijn en dat jullie dit speciale moment met mij willen en kunnen delen.

Tot slot wil ik de mensen noemen die het allerbelangrijkste zijn in mijn leven. Mijn vader, die dit helaas ook niet meer mag meemaken. Mijn moeder Corry, mijn broer Christian en mijn zoon Oscar. Jullie betekenen ongelofelijk veel voor mij.

Ik heb gezegd.

Bibliografie

Al-Haj, H., Itai, A. & Wintner, S. (2014). Lexical Representation of Multiword Expressions in Morphologically-complex Languages. *International Journal of Lexicography,* 27(2), 130-170. https://doi.org/10.1093/ijl/ect036.

Barera, M. (2020). Mind the Gap: Addressing Structural Equity and Inclusion on Wikipedia. Geraadpleegd op 29-01-2024: https://rc.library.uta.edu/uta-ir/handle/10106/29572.

Barrett, G. (2023). 'Defin-O-Bots: Challenging A.I. to Create Usable Dictionary Content.' Paper presented at the 24th Biennial Conference of the Dictionary Society of North America. Boulder, CO, USA, 31 mei – 3 juni 2023.

Bender, E. M. (2022) *Emily M. Bender: AI in reality*. Nieuwsbericht door Joyce Parvi. University of Washington, Department of Linguistics. Geraadpleegd op 20-01-2024: https://linguistics.washington.edu/news/2022/09/15/emily-m-bender-ai-reality.

Boon, C.A. den & Hendrickx, R. (2022). Dikke Van Dale Online. Van Dale, Utrecht.

Bouma, G., Odijk, J. & Tiberius, C. (2024). 'Towards a Dutch Parseme Corpus.' Poster presented at the UniDive 2nd General Meeting. Naples. 8-9 februari 2024.

Brandt Corstius, H. (1978). *Computer-taalkunde*. Muiderberg: Dick Coutinho.

Colman, L. & Tiberius, C. (2018). A Good Match: a Dutch Collocation, Idiom and Pattern Dictionary Combined. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, Ljubljana, Slovenia, 17-21 juli 2018. 233-246.

Corpas Pastor, G., & Colson, J.-P. (eds) (2020). *Computational Phraseology*. John Benjamins Publishing Company.

de Marneffe, M.C., Manning, C.D., Nivre, J. & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308. https://doi.org/10.1162/coli_a_00402.

De Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography* 36(4), 355-387. https://doi.org/10.1093/ijl/ecad021.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text & Talk*, *20*(1), 29–62. https://doi.org/10.1515/text.1.2000.20.1.29.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. MIT Press. https://doi.org/10.7551/mitpress/9780262018579.001.0001.

Ježek, E. & Hanks, P. (2010). What Lexical Sets Tell Us about Conceptual Categories. Lexis. *Journal in English Lexicology*, 4, 7-22. https://doi.org/10.4000/lexis.555.

Ježek, E., Magnini, B., Feltracco, A., Bianchini, A. & Popescu, O. (2014). T-PAS; A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 890-895. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014.

Ježek, E. (2016). *The Lexicon : An Introduction*. Oxford University Press.

Kilgarriff, A. (1997). Putting frequencies in the dictionary, *International Journal of Lexicography*, 10(2), 135–155. https://doi.org/10.1093/ijl/10.2.135.

Marini, C. & Ježek, E. (2019). CROATPAS: A Resource of Corpus-derived Typed Predicate Argument Structures for Croatian. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CliC-it)*. Bari, Italy.

Meľčuk, I. A. (1995). Phrasemes in Language and Phraseology in Linguistics. In Everaert, M. B. H., Linden, van der E.-J., Schenk, A, Schreuder, R. (eds) *Idioms : structural and psychological perspectives*. Lawrence Erlbaum. 167-232.

Meľčuk, I. A. (2023). *General Phraseology : Theory and Practice*. John Benjamins Publishing Company.

Nazar, R. & Renau, I. (2016). A Taxonomy of Spanish Nouns, a Statistical Algorithm to Generate it and its Implementation in Open Source Code. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1485–1492. https://aclanthology.org/L16-1236.

Odijk, J., Kroon, M., Baarda, T., Bonfil, B. & Spoel, S. (te verschijnen). MWE-finder: Querying for multiword expressions in large Dutch text corpora. In Giouli, V. & Barbu Mititelu, V. (eds) *Multiword expressions in lexical resources. Linguistic, Lexicographic and Computational perspectives.* (Phraseology and Multiword Expressions). Berlin: Language Science Press.

Renau, I. & Nazar, R. (2016). Automatic extraction of lexico-semantic patterns from corpora. In Margalidatze, T. & Meladze, G. (eds), *Proceedings of the XVII Euralex International Congress*. Tbilisi (Georgia), 823-830.

Sag, I.A., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In Gelbukh, A. (eds) *Computational Linguistics and Intelligent Text Processing*. CICLing 2002. Lecture Notes in Computer Science, vol 2276, 1-15. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45715-1_1.

Savary, A., Cordeiro, S. & Ramisch, C. (2019). Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 79–91, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5110.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Stubbs, M. (2002). *Words and phrases : corpus studies of lexical semantics*. Blackwell.

Stubbs, M. (2009). Memorial Article: John Sinclair (1933–2007): The Search for Units of Meaning: Sinclair on Empirical Semantics. *Applied Linguistics*, *30*(1), 115–137. https://doi.org/10.1093/applin/amn052.

Tiberius, C. & Schoonheim, T. (2014). *A Frequency Dictionary of Dutch: Core vocabulary for learners*. Routledge Frequency Dictionaries. Routledge.

Tiberius, C., Krek, S., Depuydt,K., Gantar, P., Kallas, J., Kosem, I. & Rundell, M. (2021). Towards the ELEXIS data model: defining a common vocabulary for lexicographic resources. In *Proceedings of the eLex 2021 conference*. 56-77.

Vondřička, P. (2019). Design of a Multiword Expressions Database. *The Prague Bulletin of Mathematical Linguistics*. 112. 83-101. 10.2478/pralin-2019-0003.

Corpora

British National Corpus - BNC [Dataset]. Beschikbaar: http://www.natcorp.ox.ac.uk/ [Online service]. Beschikbaar in Sketch Engine: https://www.sketchengine.eu/.

Corpus Gesproken Nederlands - CGN (Versie 2.0.3) (2014) [Dataset]. Beschikbaar bij het Instituut voor de Nederlandse Taal: http://hdl.handle.net/10032/tm-a2-k6.

Corpus Hedendaags Nederlands - CHN [Online service]. Beschikbaar bij het Instituut voor de Nederlandse Taal: http://hdl.handle.net/10032/tm-a2-s8.

PARSEME-Corpora (Versie 1.3) (2023) [Dataset]. Beschikbaar bij: LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University: http://hdl.handle.net/11372/LRT-5124.

Noten

1 https://openai.com/chatgpt
2 Corpus Hedendaags Nederlands, Het Belang van Limburg, 20-4-2021.
3 https://ivdnt.org/woordenboeken/
4 Corpus Hedendaags Nederlands, Het Nieuwsblad, 10-7-2009.
5 https://woordcombinaties.ivdnt.org
6 De definities in deze paragraaf zijn ontleend aan de *Dikke Van Dale Online*.
7 Een werk doen dat nooit af komt (*Dikke Van Dale Online*).
8 Al Haj et al. (2013:130) omschrijven 'Multiword Expressions' als 'sequences of orthographic words that for various reasons must be stored in computational lexicons as a unit'.
9 https://neurocampus.vrijetijd.nl/info?lid=2965
10 Deze varianten zijn allemaal terug te vinden in het British National Corpus geraadpleegd in Sketch Engine (28-01-2024).
11 https://gretel5.hum.uu.nl/home
12 https://commission.europa.eu/resources-partners/etranslation_en.
13 Corpus Gesproken Nederlands geraadpleegd op 29-01-2024.
14 Corpus Hedendaags Nederlands geraadpleegd op 29-01-2024.
15 Alle zoekopdrachten in het Corpus Hedendaags Nederlands zijn uitgevoerd op 29-01-2024.
16 https://lindat.mff.cuni.cz/services/udpipe/
17 https://universaldependencies.org/format.html
18 https://gitlab.com/parseme/corpora. Binnen PARSEME worden de volgende categorieën onderscheiden: LVC - light verb construction; VID - verbal idiom; IRV - inherently reflexive verb; VPC - verb particle construction; MVC - multi-verb construction.
19 https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/?page=home
20 TNE heeft duidelijke raakvlakken met andere benaderingen waarin vorm en inhoud aan elkaar gekoppeld worden, zoals constructiegrammatica en Fillmore's framesemantiek. Alhoewel deze benaderingen van oorsprong een ander vertrekpunt hebben, zien we tegenwoordig steeds meer synergieën.
21 British National Corpus: Stephanie Howard, Conspiracy of love, 1993. Geraadpleegd in Sketch Engine op 26-01-2024.
22 British National Corpus: *Today* newspaper, 1992. Geraadpleegd in Sketch Engine op 26-01-2024.
23 De ontologie is te vinden op de website van het project *Pattern Dictionary for English Verbs*: https://pdev.org.uk/#!
24 https://pdev.org.uk/#!
25 https://universaldependencies.org/
26 Bijv. UDPipe: https://lindat.mff.cuni.cz/services/udpipe/ en Stanza: https://stanfordnlp.github.io/stanza/
27 http://stanza.run/
28 https://ucrel.lancs.ac.uk/usas/
29 https://ucrel.github.io/pymusas/
30 http://www.tecling.com/verbario
31 In deze extractie wordt alleen semantische informatie gebruikt en geen syntactische.
32 ChatGPT3.5

Prof.dr. C.P.A. Tiberius

1992	Licentiaat Vertaler Frans-Russisch, Hoger Instituut voor Vertalers en Tolken, Universiteit Antwerpen
1995	Doctoraal Taal, Spraak en Informatica (cum laude), Katholieke Universiteit Nijmegen
2001	Promotie University of Brighton, Verenigd Koninkrijk (*Architectures for Multilingual Lexical Knowledge Representation)*
2000-2006	Research Fellow Surrey Morphology Group, University of Surrey, Verenigd Koninkrijk
2006-	Computerlinguïst Instituut voor de Nederlandse Taal (voorheen Instituut voor Nederlandse Lexicologie)
2023	Benoeming tot hoogleraar Computerlinguïstiek, Universiteit Leiden, Faculteit der Geesteswetenschappen, Leiden University Centre for Linguistics

Universiteit Leiden