Aspekte van die ontwerp en samestelling van ’n veeltalige aanlyn termbank vir Suid-Afrikaanse universiteitstudente

deur

Michelle Goosen

Studentenommer: 29063893

’n Verhandeling in vervulling met die vereistes vir die graad

Magister Artium Toegepaste Taalstudie

in die Departement Afrikatale by die

UNIVERSITEIT VAN PRETORIA

FAKULTEIT GEESTESWETENSKAPPE

STUDIELEIER: Prof. E. Taljard

Erkennings

Baie dankie aan Die Suid-Afrikaanse Akademie vir Wetenskap en Kuns vir die beurs wat aan my toegestaan is. Dié beurs het my finansieel in staat gestel om my studie suksesvol te voltooi.

Aan my studieleier, prof. Elsabé Taljard, baie dankie vir die voorreg dat ek hierdie verhandeling onder u leiding kon gedoen het. U bydrae is van onskatbare waarde.

Baie dankie aan die indiwidue, en in besonder aan Ryno, wat my bemoedig het en in my geglo het.

Opsomming: Hierdie studie stel ondersoek in na die daarstel van ’n veeltalige aanlyn termbank. Die doel van so ’n veeltalige aanlyn termbank sou wees om aan die meerderheid van Suid-Afrikaanse studente ’n platform in hulle sterkste taal te bied om fundamentele konsepte in verskeie vakgebiede te begryp. So ’n veeltalige aanlyn termbank sal ook die behoefte aanspreek om studente die nodige ondersteuning te bied om hulle studies in die toegekende tyd suksesvol te kan voltooi en hulle toe te rus met gevorderde akademiese geletterdheidsvlakke. Daar word ondersoek ingestel na die mate waartoe korpusgebaseerde terminologie as metodologiese aanloop tot die vestiging van ’n veeltalige aanlyn termbank aangewend kan word. Aspekte waaraan daar spesfiek aandag gegee word, is die semi-outomatiese onttrekking van: terme, verklarende inligting, kollokatiewe inligting en gebruiksvoorbeelde vanuit doelgemaakte korpora. Daar word geargumenteer dat die verklarende inligting wat semi-outomaties onttrek word gebruik kan word om terminologiese definisies saam te stel, aangesien vakspesialiste terme bewustelik/onbewustelik verduidelik wanneer vakspesifieke tekste saamgestel word. Verskillende moontlikhede word oorweeg vir die aanbieding van kollokatiewe inligting. Daar word voorgestel dat dié inligting as ’n aparte datakategorie aangebied moet word. Kollokatiewe inligting en gebruiksvoorbeelde kan nie vanuit tale wat hulpbronarm is onttrek word of net bloot vanuit die brontaal (Engels) na die teikentaal (een van die 10 amptelike Afrikatale) vertaal word nie. Die voorstel word gemaak dat die betrokke inligting deur die terminoloog, in samewerking met die vertaler en vakspesialis, in die teikentaal geskep moet word. As teoretiese raamwerk word die Kommunikatiewe Teorie van Terminologie (Eng: *Communicative Theory of Terminology*, afgekort as CTT) as uitgangspunt gebruik.

Sleutelwoorde: VEELTALIGE AANLYN TERMBANK, SEMI-OUTOMATIESE TERMIDENTIFISERING, ONTTREKKING VAN VERKLARENDE INLIGTING, GEBRUIKSVOORBEELDE, KOLLOKATIEWE INLIGTING, DOELGEMAAKTE KORPORA, KORPUSGEBASEERDE TERMINOLOGIE, SKETCH ENGINE

Abstract: This study considers the establishment of a multilingual online term bank. The aim of such a multilingual online term bank would be to create a platform for the majority of South African students where they would have access

to the fundamental concepts of different subject fields in their strongest language. Such a multilingual online term bank would address the need of students to receive the necessary support to successfully complete their studies in the given timeframe and provide them with advanced academic literacy levels. It is considered to what extend corpus-based terminology as a methodological approach can be used for the establishment of a multilingual online term bank. Aspects which are specifically addressed is the semi-automatic extraction of: terms, explanatory information, collocational information and usage examples from language for special purposes (LSP) corpora. It is argued that the explanatory information that has been extracted semi-automatically can be used to compile terminological definitions, since subject specialists explain terms consciously/unconsciously while compiling subject specific texts. Various possibilities are considered for the presentation of collocational information. It is suggested that collocational information must be presented as a separate data category. Collocational information and usage examples cannot be extracted from languages that do not have sufficient resources nor can it be translated from the source language (English) to the target languages (one of the 10 official African languages). It is suggested that collocational information and usage examples must be created in the target language by the terminologist, in co-operation with the translator and subject specialist. The theoretical framework the *Communicative Theory of Terminology* (CTT) is used as premise.

***KeyWord*:** MULTILINGUAL ONLINE TERM BANK, SEMI-AUTOMATIC TERM EXTRACTION, EXTRACTION OF EXPLANATORY INFORMATION, USAGE EXAMPLES, COLLOCATIONAL INFORMATION, LANGUAGE FOR SPECIAL PURPOSES CORPORA, CORPUS-BASED TERMINOLOGY, SKETCH ENGINE

INHOUDSOPGAWE

HOOFSTUK 1

OORSIG VAN STUDIE

1.1 Inleiding

Tersiêre opleiding, vir meeste Suid-Afrikaanse universiteitstudente, geskied deur ’n onderrigmedium, meestal Engels, wat nie hulle sterkste taal is nie. Ter illustrasie is statistiek met betrekking tot die taalprofiel van die Universiteit van Pretoria se studente ingewin. Hierdie statistiek van November 2015 het aangetoon dat uit ’n studentepopulasie van 49 152, 36% van die studente se sterkste taal ’n Afrikataal is, Afrikaans 26,8% van die studente se sterkste taal uitmaak en 29,1% van die studente se sterkste taal Engels is. Hieruit is dit duidelik dat Engels wat vir alle praktiese doeleindes die onderrigmedium is, nie die meerheid studente — 62,8% — se sterkste taal is nie.

Tydens die Nasionale Normtoetse (NNT), wat die akademiese gereedheid van ’n matriekleerling vir universiteit meet, het Yeld (2009) tot die gevolgtrekking gekom dat die meerderheid van die studente wat die universiteitswêreld betree, ondersteuning nodig het met die begrip van akademiese terme omdat hulle nie onderrig in hul sterkste taal ontvang het nie. Verder word die Toets van Akademiese Geletterheidsvlakke (TAG) deur eerstejaarstudente aan Suid-Afrikaanse universiteite afgelê om te bepaal watter tipe akademiese ondersteuning hulle nodig het, sodat hulle hul studies in die toegekende tyd suksesvol kan voltooi. Dit is belangrik om hier te noem dat die ondersteuning wat aan hierdie studente gebied word meestal nie in hul sterkste taal sal geskied nie. In haar intreerede, *The importance of being multilingual*, merk Coetzee-Van Rooy (2010) op dat die resultate van TAG nie bestempel kan word as “onbevredigend” nie, want dit is nie die ware uitbeelding van studente se taalvaardighede nie - dit fokus slegs op studente se bevoegdheid in één van die tale wat hulle ken. Sy beweer verder: “The multilingualism of these students is at best ignored, and at worst (and in all likelihood most often) perceived as a problem.”

Een moontlike manier waarop (a) die veeltaligheid van studente tot hul voordeel aangewend kan word, en (b) studente akademiese ondersteuning in hul sterkste taal kan ontvang, is die beskikbaarstelling van ’n veeltalige, aanlyn termbank. Die doel van so ’n veeltalige aanlyn termbank sou wees om aan die meerderheid van Suid-Afrikaanse studente ’n platform in hul sterkste taal te bied om fundamentele konsepte in verskeie vakgebiede te begryp en so die behoefte aanspreek om studente die nodige ondersteuning te bied sodat hulle hul studies in die toegekende tyd suksesvol kan voltooi, en toegerus kan wees met gevorderde akademiese geletterdheidsvlakke. Die ideale produk behoort ’n aanlyn termbank met gesofistikeerde soekfunksies te wees wat argumentshalwe deur middel van die verskeie Suid-Afrikaanse universiteite se leerplatforms vir gebruikers beskikbaar gestel word.

1.2 Doel van die studie

Die doel van hierdie studie is die ontwerp van ’n model vir die daarstel van so ’n termbank met die fokus hoofsaaklik op Afrikatale, gegrond op stewige teoretiese beginsels van terminologie.

Aspekte wat spesifiek in hierdie studie ondersoek word, is eerstens tot watter mate korpusgebaseerde terminologie as metodologiese aanloop tot die vestiging van ’n veeltalige, aanlyn termbank aangewend kan word. Hier word daar spesifiek aandag gegee aan verskillende metodes van termidentifisering en die voor- en nadele van elk. In die tweede plek word ondersoek ingestel na die funksie en seleksie van kollokatiewe inligting en gebruiksvoorbeelde in aanlyn termbanke.

1.3 Teoretiese raamwerk

Die studie sal binne die teoretiese raamwerk van die Kommunikatiewe Teorie van Terminologie (Eng: *Communicative Theory of Terminology*, afgekort as CTT) gedoen word. Die CTT benader terminologie vanuit ’n linguistiese perspektief en is gemik daarop om die kompleksiteit van gespesialiseerde taaleenhede, d.i. terme vanuit ’n sosiale, linguistieke en kognitiewe perspektief in ag te neem. Faber Benítez (2009:114) noem dat die CTT terminologiese eenhede as ’n “stel

bepalings" beskou wat afgelei is van hul spesifieke vakgebied, konseptuele struktuur, betekenis, leksikale en sintaktiese struktuur, valensie en die kommunikatiewe konteks van gespesialiseerde diskoers. Binne die CTT is die linguistiese benadering tot terminologie korpusgebaseerd, met kognitiewe en diskursiewe aspekte in ag genome (Cabré et. al 2012). Korpusgebaseerde terminologie gebruik vakspesifieke tekste as primêre bron van terminologiese data. Indien so 'n versameling tekste aan spesifieke vereistes voldoen, staan dit as 'n korpus bekend. Meer spesifiek dan, kan 'n korpus omskryf word as 'n groot versameling van oorspronklike tekste wat volgens spesifieke kriteria in elektroniese formaat bymekaargemaak is. Die vier belangrike kenmerke van 'n korpus wat in hierdie definisie na vore kom, is: (a) oorspronklikheid, (b) die feit dat dit in elektroniese formaat is, (c) 'n verwysing na die grootte daarvan en (d) dat dit nie 'n lukrake versameling van tekste is nie, maar dat dit volgens spesifieke kriteria saamgestel word. Elkeen van hierdie kenmerke word vervolgens kortliks bespreek. 'n Teks word as oorspronklik bestempel as dit 'n voorbeeld is van egte "lewende" taal en dit uit werklike kommunikasie tussen mense in alledaagse situasies bestaan. Binne 'n terminologiekonteks moet hier egter 'n aanpassing gemaak word – hier gaan dit nie om alledaagse kommunikasie nie, maar eerder om gespesialiseerde kommunikasie binne 'n bepaalde vakgebied. Dit sluit in kommunikasie tussen vakspesialiste, of tussen vakspesialiste en semi-spesialiste, of tussen vakspesialiste en leke. Die feit dat 'n korpus in elektroniese formaat is, beteken dat dit deur middel van 'n rekenaar geprosesseer word. Die grootte van 'n korpus word nie bepaal deur vaste reëls nie, maar hang eerder af van die doel waarvoor dit gebruik gaan word. In die geval van algemene korpora geld die beginsel van 'hoe groter, hoe beter', maar in die geval van spesiale korpora wat spesifiek vir terminologiedoeleindes saamgestel word, word 'n minimum van 'n miljoen woorde as verstek geneem. Laastens, is 'n korpus nie 'n lukrake versameling van tekste nie, maar word volgens bepaalde kriteria geselekteer om op te tree as 'n verteenwoordigende voorbeeld van 'n bepaalde taal of subsisteem van daardie taal (Bowker en Pearson 2002:9-10).

Daar word tussen verskillende tipes korpora onderskei, byvoorbeeld gesproke teenoor geskrewe korpora, sinchroniese teenoor diachroniese of historiese

korpora, en oop teenoor geslote korpora. ’n Tweedeling wat van besondere belang vir hierdie studie is, is die onderskeid tussen korpora vir spesiale doeleindes, ook genoem ‘vaktaal’ (Eng: *language for special purpose corpora*, afgekort as ‘LSP corpora’) en korpora vir algemene doeleindes (Eng: *language for general purpose corpora*, afgekort as ‘LGP corpora’). Van besondere belang vir terminologie is eersgenoemde tipe, wat Bowker en Pearson (2002:12) soos volg definieer: “It could be restricted to the LSP of a particular subject field, to a specific text type, to a particular language variety or to the language used by members of a certain demographic group (e.g. teenagers). Because of its specialised nature, such a corpus cannot be used to make observations about language in general” (Bowker en Pearson 2002:12).

Met betrekking tot die gebruik van elektroniese korpora in algemene leksikografie, beskou Taljard (2004:174) dit as ’n gevestigde praktyk in Suid-Afrika, aangesien dat die nege Nasionale Leksikografie Eenhede van die Suid-Afrikaanse Afrikatale almal tot ’n mindere of meerdere mate gebruik maak van korpora vir die samestelling van hul verskeie algemene woordeboeke. Sy voeg by dat die gebruik van korpora vir terminologiese of spesiale doeleindes ’n mindere algemene praktyk is, met die gebruik van korpora wat tot algemene leksikografie beperk word. Die studie sal poog om die fondasies te lê vir die gebruik van korpora vir vaktaaldoeleindes in Suid-Afrika met die oog daarop dat dit ’n algemene praktyk sal word.

Volgens Faber Benítez (2009:115) het die CTT wel ’n paar tekortkominge. Sy is van mening dat daar binne die betrokke teorie meer lig gewerp moet word op konseptuele semantiek en semantiese betekenis. Met dié inligting in ag genome, weeg die voordele van die CTT nog steeds swaarder op teen die tekortkominge van die teorie.

1.4 Termidentifisering

Een van die belangrikste take van ’n terminoloog in enige terminologiese aktiwiteit is die identifisering van terme wat tot ’n bepaalde vakgebied behoort. Vir die doel van hierdie studie word ’n onderskeid getref tussen die tradisionele en moderne benaderings tot termidentifisering.

Die tradisionele benadering maak staat op die kennis van ’n vakspesialis wat die tekste moet lees, die inhoud moet verstaan en die potensiële terme handmatig moet ekserpeer. Dit is nie altyd haalbaar om hierdie metode te volg nie nie, aangesien tekste - in die meeste gevalle - te groot is (Warburton 2014:1). Dit is verder ’n subjektiewe benadering, aangesien kundiges verskillende opinies mag hê oor die relevansie van die geselekteerde data en dus bestaan die moontlikheid dat potensiële terme uitgelaat word. Dit is ook ’n tydrowende en duur proses.

Die moderne, korpusgebaseerde benadering tot termidentifisering berus op die semi-outomatiese, rekenaarmatige onttrekking van potensiële terme uit ’n korpus. Een van die voordele van korpusgebaseerde termonttrekking is dat dit die terminoloog in staat stel om enorme hoeveelhede data in ’n ommesientjie te prosesseer. Nchabeleng (2011:4) noem dat die metode “semi-outomaties” genoem word, omdat die rekenaar die terme nie ten volle outomaties kan onttrek nie. “What the computer software initially extracts are only term candidates. Candidates may not all be terms.” Menslike intervensie is steeds nodig om die “geraas” (ongeldige potensiële terme) en “stilte” (ontbrekende geldige potensiële terme) wat in termlyste mag voorkom, onderskeidelik te verwyder en by te voeg. De Schryver en Taljard (2002:46) argumenteer dat in ál die benaderinge tot semi-outomatiese termonttrekking, die mens - verkieslik ’n vakspesialis - die finale seggenskap het in die besluit of die potensiële terme wat deur die sagteware geïdentifiseer is, geldige terme is. Die lys van kandidaatterme bevat ook ander nuttige terminologiese data, onder andere frekwensiedata en data oor die verspreiding van ’n spesifieke term oor verskillende tekste. Aangesien hierdie proses rekenaarmatig uitgevoer word, skakel dit die subjektiewe, menslike element tot ’n groot mate uit. Die proses is vinniger en meer akkuraat as die tradisionele benadering. Warburton (2014:2) is van mening dat (semi-outomatiese) termonttrekking noodsaaklik is vir die ontwikkeling van ’n terminologiese databasis.

Volgens Bowker en Pearson (2002:165) kan die sagteware wat tans vir termonttrekking beskikbaar is, eentalige, sowel as veeltalige funksies verrig. Sagteware vir eentalige termonttrekking ontleed eentalige spesiale korpora om

potensiële terme te identifiseer. Aan die ander kant word sagteware vir veeltalige termidentifisering gebruik om veeltalige, meestal parallelle korpora te ontleed om potensiële terme en hul vertaalekwivalente te identifiseer. Binne die studie sal die gebruik van sagteware vir eentalige termidentifisering ondersoek word, omdat termidentifisering meestal uit eentalige Engelse korpora gedoen word.

Daar is tans verskeie sagteware vir semi-outomatiese termonttrekking beskikbaar. Binne die studie word die sagteware *Sketch Engine* (https://www.sketchengine.co.uk) geselekteer vir die proses. *Sketch Engine* word gebruik om vas te stel hoe woorde in ’n teks optree. Die sagteware werk aanlyn en bied aan die gebruiker ’n verskeidenheid korpora in verskillende tale. Met die betrokke sagteware kan beide enkelwoord- en meerwoordterme onttrek word.

1.5 Gebruiksvoorbeelde in aanlyn termbanke

Bowker en Pearson (2002:16) gaan van die veronderstelling uit dat vaktaalgebruikers daarvan bewus moet wees hoe terme gebruik word. “In addition to information about what a term means, they also need information about how to use that term in a sentence. This information can be provided by presenting the terms in context instead of isolation.” Deur die vaktaalgebruiker toe te rus met gebruiksvoorbeelde word die inligting in ’n definisie aangevul, dit dui aan hoe die term in konteks gebruik word, dit demonstreer tipiese kollokasies en dit dui die gepaste register aan. Dit is dus in die beste belang van die gebruiker om hom/haar toe te rus met ’n term en definisie, verder uitgebrei met ’n gebruiksvoorbeeld, sodat hy/sy die term beter kan begryp.

Met betrekking tot die insluit van gebruiksvoorbeelde in Suid-Afrikaanse (algemene) woordeboeke, het Hiles (2009) ’n ondersoek gedoen na die rol van gebruiksvoorbeelde in huidige Suid-Afrikaanse skoolwoordeboeke, onder andere om te bepaal hoe dit verbeter kan word. Sy het ’n lukrake keuse van vyf Suid-Afrikaanse woordeboeke gemaak - van hierdie woordeboeke is gebaseer op woordeboeke wat in die Verenigde Koninkryk en Verenigde State gepubliseer is, en die ander is in Suid-Afrika saamgestel. Sy maak ’n terminologiese onderskeid tussen gebruiksvoorbeelde in die vorm van frases, wat sy dooie

gebruiksvoorbeelde noem, en dié wat volledige sinne is, waarna sy as lewende gebruiksvoorbeelde verwys. Sy het bevind dat, hoewel meeste van die gebruiksvoorbeelde - hetsy frases of sinne - ondersteuning in een of ander vorm bied, is hierdie gebruiksvoorbeelde nie op die mees effektiewe manier geskryf of gekies nie. Met sommige van die gebruiksvoorbeelde moet gebruikers deur lang sinne swoeg om een of ander vorm van ondersteuning te vind, en in ander gevalle is die gebruiksvoorbeelde te kort om enige ondersteuning te bied. Sy het ook vasgestel dat grammatiese ondersteuning veral belangrik is in tweetalige woordeboeke - ofskoon die gebruiker die woord sal ken deur middel van die vertaalekwivalent, het die gebruiker hulp nodig met die grammatiese struktuur van ’n taal wat nie sy/haar sterkste taal is nie. Dus ontstaan die behoefte binne die Suid-Afrikaanse konteks, om (a) gebruiksvoorbeelde in die vorm van sinne in plaas van frases aan gebruikers te voorsien, aangesien sinne meer ondersteuning bied as frases (Hiles 2009:104); (b) gebruiksvoorbeelde te selekteer wat nie te lank of te kort is nie, omdat dit die gebruiker kan verwar, laat moed opgee of as nutteloos beskou kan word; en (c) waar moontlik, om grammatiese ondersteuning soos kollokasies in te sluit, omdat dit die gebruiker op ’n konseptuele, sowel as ’n gebruiksvlak help (Taljard 2016:553).

Atkins en Rundell (2008:458) noem dat ’n goeie gebruiksvoorbeeld natuurlik en kenmerkend, informatief en verstaanbaar moet wees. Elkeen van dié kenmerke word kortliks bespreek. ’n Gebruiksvoorbeeld is natuurlik en kenmerkend wanneer dit ’n uiting is wat die gebruiker waarskynlik sal hoor by ’n moedertaalspreker van die betrokke taal. ’n Gebruiksvoorbeeld is informatief wanneer dit die definisie aanvul en die gebruiker help om die term beter te verstaan. En laastens, ’n gebruiksvoorbeeld word as verstaanbaar geklassifiseer wanneer sinnelose, moeilike leksis en sinstrukture vermy word. Aan die ander kant is Hiles (2009:27) van mening dat gebruiksvoorbeelde wat verwarrend, onvanpas, of kwetsend is, die beleid van die woordeboeksamesteller verontagsaam of die onreëlmatige gebruik van die woord illustreer, as swak gebruiksvoorbeelde beskou word. Die onus rus dus op die leksikograaf om gebruiksvoorbeelde te selekteer wat aan al hierdie vereistes voldoen.

Fuertes-Olivera en Arribas-Baño (2008:129) is van mening dat daar drie benaderings tot gebruiksvoorbeelde in moderne leksikografie bestaan, naamlik: gebruiksvoorbeelde wat deur ’n leksikograaf geskep is (sogenaamde studeerkamervoorbeelde), gebruiksvoorbeelde wat vanuit ’n korpus onttrek is, en gebruiksvoorbeelde wat vanuit ’n korpus onttrek is en tot ’n mindere of meerdere mate aangepas is deur die leksikograaf op grond van die kennis van sy/haar moedertaal. Met die beskikbaarheid van korpora het die fokus verskuif vanaf die gebruik van studeerkamervoorbeelde na die gebruik van korpusgebaseerde voorbeelde, omdat studeerkamervoorbeelde eerder die betekenis van ’n term verduidelik, in plaas van om die gebruik daarvan aan te dui. Landau (2001:210) beweer in dié verband: “Using invented examples is like fixing a horse race: the lexicographer invents an example to justify his definition, instead of devising a definition to fit the examples.” Hierdie werkswyse is dus nie ’n ware weerspieëling van egte “lewende” taal nie. Kundiges is ten gunste van die gebruik van korpusgebaseerde voorbeelde, omdat dit ’n werklikheidsgetroue voorbeeld van taalgebruik is (Fuertes-Olivera en Arribas-Baño 2008:129). Rundell (1998:334-335) vat dit soos volg saam: “The corpus provides natural and typical examples that clearly illustrate the points that need to be made, there is no conceivable reason for not using them.” Binne die studie word gebruiksvoorbeelde wat vanuit spesiale korpora onttrek is, ondersoek. Wanneer gebruiksvoorbeelde vir terminologiese doeleindes geselekteer word, plaas dit ’n addisionele las op die terminoloog / leksikograaf: ten einde ’n oordeel te vel of ’n gebruiksvoorbeeld wel aan Atkins en Rundell (op cit.) se vereistes voldoen, is ’n basiese kennis van die betrokke vakgebied ook nodig.

Volgens Humblé (2001:61-62) moet daar onderskeid getref word tussen gebruiksvoorbeelde vir dekodering en gebruiksvoorbeelde vir enkodering. “A decoding learner is interested in the meaning of a lexical item, whereas an encoding learner is interested in a word’s syntactic features and collocates.” Binne die studie word geargumenteer dat gebruiksvoorbeelde binne die konteks van ’n veeltalige aanlyn termbank vir Suid-Afrikaanse studente noodsaaklik is, en verskillende metodes vir die identifisering van goeie gebruiksvoorbeelde word ondersoek. Dit is ook volgens Frankenberg-Garcia (2012:273) ’n

onderskeidende kenmerk van korpusgebaseerde woordeboeke dat die meeste inskrywings gebruiksvoorbeelde of frases bevat wat vanuit korpora gekopieer of aangepas is. ’n Goed gekose korpusgebaseerde gebruiksvoorbeeld, met kollokasies in ag genome, kan beide die funksies van dekodering en enkodering vervul. Maar met die oog daarop om ’n veeltalige aanlyn termbank op te rig wat vakspesifieke glossariums bevat wat deur universiteitstudente - die teikengroep - geraadpleeg word, sal die primêre funksie wees om gebruiksvoorbeelde vir dekoderingsdoeleindes te gebruik, omdat die teikengroep hoofsaaklik in die betekenis van ’n term belangstel.

In ’n studie wat deur Frankenberg-Garcia (2012:287-289) uitgevoer is om die doeltreffendheid van korpusgebaseerde gebruiksvoorbeelde te bepaal, het sy bevind dat ’n enkele korpusgebaseerde gebruiksvoorbeeld studente help om terme korrek te gebruik (eerder as om net ’n definisie tot hul beskikking te hê) en dat veelvuldige gebruiksvoorbeelde per term nog meer begripsvoordele inhou.

Gebruiksvoorbeelde word nie in die meeste vakwoordeboeke weergegee nie, waarskynlik as gevolg van spasiebeperkinge. Die gedrukte woordeboeke wat wel gebruiksvoorbeelde bevat, het slegs spasie vir enkele gebruiksvoorbeelde (Bowker en Pearson 2002:16). In ’n aanlyn termbank verval die eis om ruimtebesparing en is die insluiting van een of meer gebruiksvoorbeelde dus nie problematies nie. Dit is dus haalbaar om ’n goed deurdagte en goed gekose korpusvoorbeeld aan vaktaalgebruikers te voorsien wat vir dekoderingdoeleindes gebruik gaan word.

HOOFSTUK 2

TEORETIESE RAAMWERK: DIE KOMMUNIKATIEWE TEORIE VAN TERMINOLOGIE (CTT)

2.1 Inleiding

In hierdie hoofstuk word ’n kontekstualisering van die Algemene Teorie van Terminologie gegee (Eng: *General Theory of Terminology*, afgekort as GTT) wat ’n blik gee op terminologie as ’n dissipline. Daar word gekyk na die aspekte van ’n teorie en wat dit behels om ’n toereikende teorie van terminologie saam te stel. Die teoretiese raamwerk, die Kommunikatiewe Teorie van Terminologie (Eng: *Communicative Theory of Terminology*, afgekort as CTT), waarbinne die onderhawige studie geskied, word bespreek. Laastens word daar ook gekyk na die samestelling, doel en funksie van doelgemaakte korpora binne die CTT model.

2.2 Historiese kontekstualisering

Terminologie - die dissipline wat gemoeid is met die bestudering en samestelling van vakterme - is in die 1930’s as ’n dissipline gevestig, danksy Eugen Wüster. Sy siening van terminologie is hoofsaaklik gebaseer op sy ervaring as ’n ingenieur wat betrokke was by die standaardisering van nasionale en internasionale terminologie wat hy as noodsaaklik beskou het vir effektiewe kommunikasie. Hy het die Tradisionele Teorie van Terminologie, nou bekend as die Algemene Teorie van Terminologie (Eng: *General Theory of Terminology*, afgekort as GTT), ontwikkel op grond van sy terminografiese werk deur die samestelling van *The Machine Tool, an Interlingual Dictionary of Basic Concepts* (Wüster 1968),’n sistematies-georganiseerde Frans-Engelse woordeboek wat gestandaardiseerde terme bevat (met ’n Duitse aanvulling), met die voorneme om te dien as ’n model vir toekomstige vakwoordeboeke. Deur sy werk het Wüster gepoog om:

- Terminologie as ’n dissipline te vestig en dit tot ’n onafhanklike wetenskap uit te bou.
- Dubbelsinnigheid in vaktale uit te skakel, d.m.v. gestandaardiseerde terminologie sodat hulle effektiewe middele van kommunikasie kan wees.

- Alle gebruikers van tegniese tale te oortuig van die voordele verbonde aan gestandaardiseerde terminologie.

Ten einde hierdie doelwitte te bereik het Wüster hom op die volgende take toegelê (Cabré 2003:165):

- Die ontwikkeling van gestandaardiseerde internasionale beginsels vir die beskrywing en dokumentering van terme.
- Die formulering van algemene beginsels van terminologie as onafhanklike dissipline.
- Die vestiging van ’n internasionale sentrum vir die versameling, verspreiding en koördinasie van inligting oor terminologie. Hierdie sentrale versamelpunt staan vandag as *Infoterm* bekend.

Die model is later verder ontwikkel deur Wüster se opvolgers wat sy idees aangevul en uitgebrei het. Hul bydraes tot die GTT sluit die volgende in (Cabré 2003:167-168):

- Die doelwit van internasionale standaardisering word uitgebrei deur voorstelle dat terminologiese ontwikkeling deel van taalbeplanning moet wees.
- Gekontroleerde sinonimie word erken.
- Sinonimie word tot ’n sekere mate aanvaar. Daar word wel aanbeveel dat dit vermy moet word in gevalle waar die doel is om die terminologie te standaardiseer.
- Fraseleer word bygevoeg tot die studie van terminologiese eenhede.
- Die betekenis van gesproke vorme word erken binne die konteks van taalbeplanning.
- Die beskrywing van die prosesse van termskepping is bygevoeg.
- Die model word aangevul deur die voorstelling van niehiërargies geordende konseptuele strukture.

Ten spyte hiervan het die GTT ’n aantal beperkinge, o.a. met betrekking tot die verhouding tussen konsepte en terme: dit neem nie die kognitiewe, linguistiese, kommunikatiewe en ander aspekte wat relevant is tot terminologie, in ag nie (Packeiser 2009:8). Die model hou verder nie rekening met die sintaksis en

pragmatiek van gespesialiseerde taal nie - dit word dus as irrelevant beskou (Faber Benìtez 2009:111) en terme word as geïsoleerde taaltekens beskou. In hierdie opsig kon die GTT nie effektief vir vertalingdoeleindes gebruik word nie. Wüster (en sy opvolgers) se benadering tot terminologie het die diachroniese aspek van terme uitgesluit, en terminologie was uitsluitlik sinchronies. Dit is duidelik dat die GTT daarna gestrewe het om die vormlike aspekte van vaktale aan te spreek, en kan dus nie rekening gee van die spesifiekheid van semantiese aspekte van gespesialiseerde tekens nie.

Cabré (2003:164) is van mening dat daar ’n behoefte is vir: (a) die ontwikkeling van ’n teorie van terminologie wat handel oor terminologiese eenhede, i.p.v. terminologie, en (b) ’n teorie wat rekenskap kan gee van die kompleksiteit van terme in hul egte en gevarieerde kommunikatiewe konteks. Hierdie is ’n belangrike afwyking van die Wüsteriaanse benadering waarvolgens terme in isolasie bestudeer is. ’n Terminologiese eenheid word as die mees komplekse eenheid beskou, omdat dit dieselfde meerdimensionaliteit vertoon as terminologie as ’n dissipline. ’n Terminologiese eenheid word verryk deur die teorie van kennis, die teorie van kommunikasie en die teorie van taal. Wanneer ’n teorie van terminologie saamgestel word, is een van die belangrike vrae wat gevra moet word: Wat is ’n teorie en wat beteken dit om ’n teorie van terminologie saam te stel? Binne die positivistiese benadering tot kennis is ’n teorie ’n sisteem van voorstellings wat afgelei is van ’n klein aantal beginsels. Enige teorie behoort sover as moontlik in eenvoudige, volledige en presiese vorm voorgestel te word as ’n reeks eksperimentele wette. Die voorwaardes van eenvoudigheid, volledigheid en akkuraatheid baan die weg na ’n logies-formele analise. Vanuit hierdie perspektief word ’n teorie as ’n stel hipoteses beskou wat, wanneer dit gevestig is, gestaaf of weerlê moet kan word (Cabré 2003:179-180). Cabré voeg by dat ’n teorie verskillende vlakke van toereikendheid kan hê. “A theory is observationally adequate if it permits the description of the observed data. It is descriptively adequate if, besides permitting the description of the observed data, it permits the description of the non-observed ones which might arise. That makes it predictive. A theory is explanatory adequate if, besides being observationally and descriptively adequate, it explains how and why these data are produced and how they are obtained.” Daar kan verskeie redes wees vir die

ontwikkeling van ’n (nuwe) teorie. ’n Teorie kan naamlik ontwikkel word na aanleiding van die ontoereikendheid van ’n ander teorie, of dit kan ontwikkel word op grond van ’n intuïsie en ’n spekulatiewe proses wat lei tot hipoteses wat op hul beurt weerlê of bevestig moet word. Soms word die basiese beginsels van ’n teorie nie bevraagteken nie, maar word sekere aspekte wat voorheen nie beklemtoon is nie, uitgebrei. Dit mag ook wees dat ’n gegewe teorie ’n ander teorie kan opvolg en vervang.

Teorieë van terminologie kan as voorskriftelik of as beskrywend geklassifiseer word. Aangesien die GTT die eerste teorie van terminologie is, word dié teorie as voorskriftelik geklassifiseer. Daaropvolgende teorieë wat ontwikkel is in reaksie tot die GTT, is beskrywend. Hierdie teorieë demonstreer ’n oorhoofse neiging om die kognitiewe aspekte van (vak)taal te inkorporeer, aangesien hulle op die sosiale, kommunikatiewe en kognitiewe aspekte van terminologie fokus. “The vision they offer is more realistic because they analyse terms as they actually are used and behave in texts. One might say that these new series are representative of a *cognitive shift* in terminology” (Faber Benítez 2009:111). Die verwysing na die analise van terme soos dit binne die werklike konteks van (vak)taalgebruik voorkom, bevestig dus die gedagte van ’n klemverskuiwing van die bestudering van terme in isolasie na ’n benadering waar hierdie linguistiese eenhede binne konteks beskou word. Dit is ook die benadering wat in hierdie studie gevolg word, en hou direk verband met die korpusgebaseerde metodologie wat gevolg word.

2.3 Die Kommunikatiewe Teorie van Terminologie (CTT) as reaksie op die GTT

Soos reeds in paragraaf 1.3 genoem, geskied hierdie studie binne die teoretiese raamwerk van die Kommunikatiewe Teorie van Terminologie (Eng: *Communicative Theory of Terminology*, afgekort as CTT) (die teorie wat voorspruit uit die ontoereikenheid van die GTT). Die CTT word beskou as die beste kandidaat – wat ’n lewensvatbare, werkende teorie van terminologie is – om die GTT te vervang. “It has led to a valuable body of research on different aspects of Terminology such as conceptual relations, terminological variation, term extraction, and the application of different linguistic models to terminology.

This has helped terminology as a field … and begin to question GTT premises, which previously were not open to doubt" (Faber Benítez 2009:114-115). Voorts is hierdie teoretiese model ook geselekteer omdat dit beter toegerus is as die GTT om rekenskap te gee van verskeie scenarios wat verband hou met tegniese en wetenskaplike kommunikasie. Binne hierdie teorie van terminologie word terminologie vir die volgende doeleindes aangewend: (a) kommunikasie binne vakspesifieke gebiede, (b) kommunikasie deur tussengangers, en (c) die samestelling van vakspesifieke glossariums en woordeboeke. Binne die studie vind kommunikasie doeltreffend plaas binne vakspesifieke gebiede, d.m.v. vakspesifieke glossariums wat saamgestel is deur terminoloë (met behulp van vakspesialiste) vir semi-spesialiste (studente), sodat dit akademiese ondersteuning in hul sterkste taal bied.

Soos Sageder (2010:129) meld, is Cabré se vertrekpunt twee veronderstellings, nl.: (a) dat terminologie tegelykertyd 'n versameling van behoeftes, 'n versameling van gebruike om hierdie behoeftes aan te spreek en 'n saambindende teorie van kennis is, en (b) dat die elemente van terminologie terminologiese eenhede is. Cabré (2003:186) stel dan die volgende vraag: Hoe kan 'n teorie geformuleer word wat die verskillende aspekte van terminologie inkorporeer? In reaksie tot hierdie vraag stel sy binne die CTT die metafoor *die teorie van deure* (Eng: *the theory of doors*) voor wat die begrip, analisering en toegang tot terminologiese eenhede, verteenwoordig. Hierdie model het ten doel om die sentrale objek, die terminologiese eenheid, aan te spreek. Die voorneme is dus om rekening te gee van terminologiese eenhede vanuit 'n teorie van natuurlike taal. So 'n model moet ook aan sekere voorwaardes voldoen sodat daar onderskeid getref kan word tussen taaleenhede met dieselfde struktuur, nl.: woorde aan die een kant, en eenhede wat gespesialiseerde kennis uitdruk, bv. gespesialiseerde fraseologiese-, morfologiese- of sinseenhede aan die ander kant. Cabré vergelyk 'n terminologiese eenheid met 'n veelvlak waar dit vanuit drie perspektiewe benader kan word, nl.: 'n kognitiewe perspektief (die konsep), 'n linguistiese perspektief (die term), en 'n kommunikatiewe perspektief (die situasie). Elkeen van hierdie perspektiewe word vervolgens bespreek (Cabré 2003:184). Binne die kognitiewe perspektief voldoen die terminologiese eenheid aan die volgende voorwaardes:

- Dit maak staat op ’n tematiese konteks.
- Dit neem ’n baie spesifieke posisie binne die konseptuele struktuur in.
- Spesifieke betekenis word bepaal deur die terminologiese eenheid se posisie in die konseptuele struktuur.
- Die betekenis is eksplisiet en word beskou as ’n eienskap van die eenheid.
- Dit word erken en gedissemineer deur deskundiges.

Binne die linguistiese perspektief voldoen die terminologiese eenheid aan die volgende voorwaardes:

- Hulle is leksikale eenhede óf deur hul leksikale oorsprong óf deur die proses van leksikalisering.
- Hulle kan ’n leksikale of sintaktiese struktuur hê.
- As ’n leksikale struktuur benut die terminologiese eenhede alle hulpmiddels m.b.t. woordvorming en die prosesse vir die verkryging van nuwe eenhede.
- Hulle mag vormlik ooreenstem met eenhede wat tot algemene diskoers behoort.
- Hulle kom voor as naamwoorde, werkwoorde, adjektiewe of bywoorde of nominale, verbale, adjektiwiese of bywoordelike strukture.
- Hulle behoort tot een van die breë semantiese kategorieë, nl.: entiteite, gebeurtenisse, eienskappe of betrekkinge.
- Betekenis binne ’n spesiale vakgebied is diskreet.

Binne die kommunikatiewe perspektief voldoen die terminologiese eenheid aan die volgende voorwaardes:

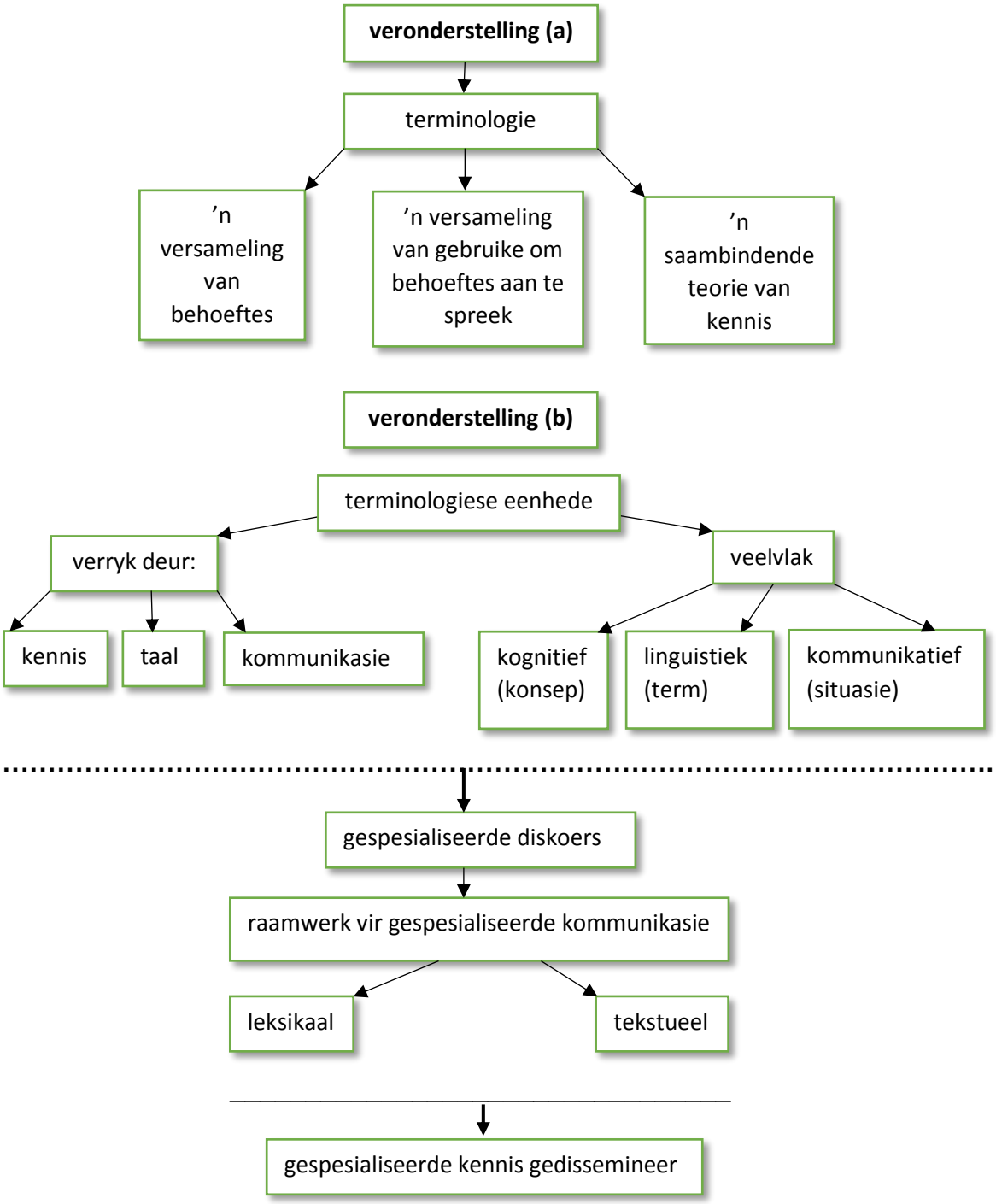
- Hulle kom voor in gespesialiseerde diskoers.
- Hulle pas aan by die spesiale diskoers volgens hul tematiese en funksionele eienskappe.
- Hulle deel gespesialiseerde diskoerse met eenhede wat tot ander ikoniese of simboliese sisteme behoort.
- Hulle word geïnternaliseer deur ’n leerproses en word om hierdie rede deur vakspesialiste gebruik.

- Hulle word beskou as denotatief en sluit nie konnotasies uit nie.

Daar moet ook in gedagte gehou word dat, hoewel hierdie drie perspektiewe onlosmaaklik is van die terminologiese eenheid, bied elkeen van hierdie perspektiewe toegang tot die objek en die keuse van fokus bring nie mee dat die ander twee perspektiewe verwerp word nie. “According to Cabré, the CTT approaches units through the language door, but always within the general context of specialized communication” (Faber Benítez 2009:114).

Binne die raamwerk van gespesialiseerde kommunikasie word gespesialiseerde kennis (gesproke of geskrewe) gedissemineer; dit sluit die volgende in: kommunikasie tussen vakspesialiste, kommunikasie tussen vakspesialiste en semi-spesialiste, kommunikasie tussen vakspesialiste en leerders, en laastens, die popularisasie van wetenskap en tegnologie. Vir die doel van die onderhawige studie word aanvaar dat kommunikasie plaasvind tussen (a) vakspesialiste, omdat die inhoud van die vakspesifieke glossariums deur vakspesialiste saamgestel en geverifieer word om te verseker dat die korrekte gespesialiseerde kennis gedissemineer word, asook tussen (b) vakspesialiste en semi-spesialiste, omdat die vakspesifieke glossariums deur semi-spesialiste (d.i. studente) geraadpleeg word.

Die voorafgaande paragrawe word vervolgens skematies saamgevat.



Figuur 2.1. Die voorstelling van die CTT-model.

2.4 Korpora en die CTT model

Soos hierbo verduidelik, is die CTT 'n benadering waar terminologiese eenhede tegelykertyd as eenhede van taal, sosiale funksie en kognisie bestudeer word. Faber Benítez (2009:115) is van mening dat die CTT nie 'n keuse uitoefen m.b.t.

’n spesifieke linguistiese model nie, maar soos reeds in paragraaf 1.3 gemeld, is die linguistiese benadering tot terminologie binne hierdie teoretiese model, korpusgebaseerd. “… the TCT [sic] is not only interested in prescriptive terminology, i.e. the terminology established by standards or found in official databases but also (and particularly) in those terms that are actually used in L.S.P. … corpora by field experts” (Cabré et. al 2012). Hulle voeg by dat die CTT nie net die *in vitro* benadering aanneem nie, maar ook belangstel in terminologiese eenhede *in vivo*. By implikasie beteken ’n korpusgebaseerde benadering dat terme in hul natuurlike vorm in vaktaaltekste bestudeer word, met die doel om gespesialiseerde kennis te dissemineer. Gesien in die lig van die sentrale rol wat korpora binne die CTT speel, word ’n uitvoerige bespreking van die samestelling, doel en funksie van doelgemaakte korpora in die hieropvolgende paragrawe gegee.

Die vier belangrike kenmerke van ’n doelgemaakte vaktaalkorpus korpus (Eng: *special purpose corpus*) is in paragraaf 1.3 bespreek.

Pearson (1998:1) lys drie moontlike redes waarom korpora vir terminologiese of spesiale doeleindes op internasionale vlak ’n minder algemene praktyk is: (a) gebrek aan geskikte korpora, (b) ’n ongewilligheid van tradisionele terminoloë om outentieke tekste as primêre databronne te gebruik, en laastens, (c) die opvatting dat woorde fundamenteel van terme verskil en dat slegs vakspesialiste in staat is om terme te definieer (en te identifiseer). Met betrekking tot Suid-Afrika staan korpusgebaseerde terminologie nog in sy kinderskoene, hoofsaaklik a.g.v. ’n gebrek aan kundigheid t.o.v. die gebruik van korpora vir terminologiese doeleindes – dit sluit in die saamstel van korpora vir spesiale doeleindes. Om hierdie kwessie aan te spreek, kan die volgende dien as algemene riglyne wanneer ’n doelgemaakte korpus saamgestel word (Bowker en Pearson, 2002:54):

- Grootte: enige iets tussen ’n paar duisend en ’n paar honderd duisend woorde blyk nuttig te wees vir die studie van ’n vaktaal. Die nuutste navorsing dui egter daarop dat een miljoen woorde in werklikheid die minimum is.

- Volledige teks of uittreksel: volledige tekste word oor die algemeen bo uittreksels verkies, omdat belangrike inligting in volledige tekste gevind kan word.
- Aantal tekste: daar word aanbeveel dat tekste van ’n verskeidend van outeurs versamel word en nie net van een outeur nie; dit bied ’n beter oorsig van die terme binne ’n vaktaal.
- Medium: korpora wat geskrewe data bevat kan makliker saamgestel word.
- Tekstipes: wanneer die vaktaal van ’n bepaalde vakgebied bestudeer word moet slegs vakspesifieke tekste by die korpus ingesluit word.
- Outeurskap: tekste van geloofwaardige outeurs moet by die korpus ingesluit word; dit is hoogs waarskynlik dat dit meer outentieke voorbeelde en ander terminologies-relevante data bevat.
- Taal: wanneer oorspronklike taal-materiaal in ’n eentalige korpus gebruik word, sal dit meer outentieke voorbeelde van die vaktaal oplewer; parallelle korpora bestaan uit oorspronklike en vertaalde tekste.
- Verskyningsdatum: vaktaalstudies is gemoeid met die huidige stand van ’n taal en vakgebied, daarom moet die korpus uit tekste bestaan wat onlangs gepubliseer is.

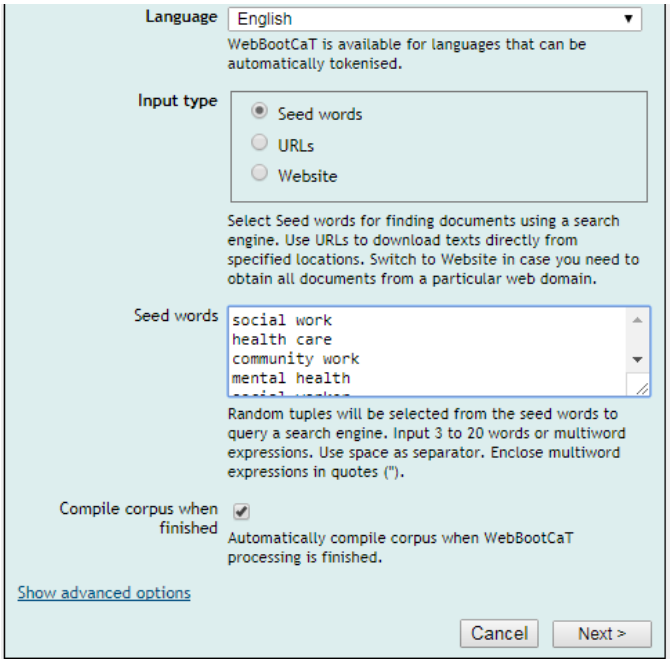
Cabré et. al 2012 wys daarop dat dit ideaal sou wees indien terminoloë eerstehandse ondervinding het van die vakgebied waarin hulle besig is; hulle behoort vertroud te wees met die tipe taal wat gebruik word en sou dit nodig wees, om inligting wat uit die korpus verkry word, aan te vul deur konsultasie met vakspesialiste. Benewens kennis van die spesifieke register wat vir kommunikasie in ’n spesifieke vakgebied gebruik word, is dit ook wenslik dat die terminoloog oor ’n sekere mate van kennis oor die inhoud van die vakgebied beskik ten einde te verseker dat akkurate data uit die korpus ingewin word. In praktyk is dit egter nie altyd moontlik nie, omdat terminoloë dikwels in ’n verskeidenheid vakgebiede aktief is. Verder word hierdie aspek ondervang deur die feit dat die tekste wat gebruik word om korpora vir spesiale doeleindes te bou, deur vakspesialiste aanbeveel word en in sommige gevalle word dit deur die outeurs self voorsien; dit verseker dat die inhoud eg en relevant is.

Wanneer ’n korpus vir spesiale doeleindes gebou word, kan tekste op verskillende maniere versamel word. In die eerste plek kan tekste wat reeds in elektroniese, d.w.s. masjienleesbare formaat is, gebruik word. Tweedens moet tekste wat slegs in harde kopie bestaan, byvoorbeeld handboeke, geskandeer word om dit in elektroniese formaat beskikbaar te stel. In die derde plek kan tekste semi-outomaties versamel word, maar nog verder, vaktaalkorpora kan semi-outomaties saamgestel word deur van tekste op die internet gebruik te maak en sagteware soos BootCaT (http://bootcat.dipintra.it/) aan te wend vir korpuskompilasie. Elkeen van hierdie versamelingstegnieke word vervolgens bespreek.

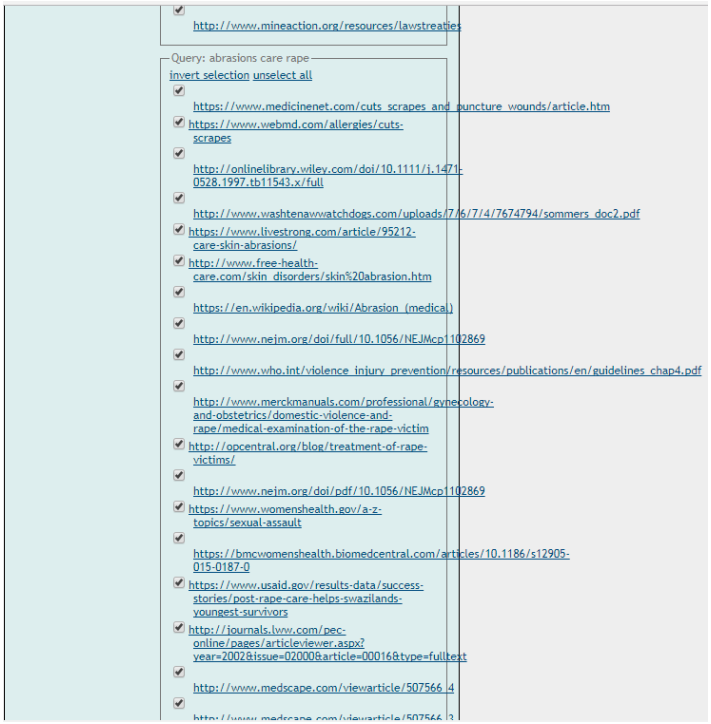
Vir die doel van die onderhawige projek sluit elektroniese tekste studiegidse, klasaantekeninge, opdragte en toets- en eksamenvraestelle in. Wanneer tekste slegs in harde kopie en nie elektronies beskikbaar is nie, word die korpusbouproses vertraag deur optiese karakterherkenningskandering (Eng: *Optical Character Recognition*, afgekort as OCR) en daaropvolgende menslike intervensie is onvermydelik om voorspelbare én onvoorspelbare skanderingsfoute te korrigeer. Wat die semi-outomatiese kompilasie van LSP korpora betref, het die beskikbaarheid van programmatuur soos byvoorbeeld BootCat baie daartoe bygedra dat sodanige korpora relatief maklik en vinnig saamgestel kan word.

BootCaT word gebruik om LSP korpora vanaf die web te *bootstrap*. Dit is die iteratiewe proses van opstapeling van terme op ’n bestaande termlys. Die korpussamesteller hoef slegs saadwoorde (Eng: *seeds*) - ’n minimum van vyf woorde word as verstek gebruik - in te sleutel wat heel waarskynlik in ’n bepaalde vakgebied sal voorkom. Ter illustrasie figuur 2.2 waar die saadwoorde van die vakgebied Maatskaplike Werk binne die betrokke sagteware ingesleutel is. Dit is belangrik dat die korpussamesteller oor voldoende vakspesifieke kennis beskik om relevante terme te lys. Nog ’n nuttige tegniek om die soektog aan te vul is om gebruik te maak van ’n bestaande vakspesifieke termlys (wat volgens die hoogste frekwensie gerangskik is) om te dien as saadwoorde. Voorts oefen die korpussamesteller die keuse uit oor die aantal URL’s wat hy/sy in die korpus wil hê (sien figuur 2.3); dit speel dus ’n bepalende faktor in die grootte van die

korpus. Die uiteindelike kwaliteit van die inhoud van die korpus berus op menslike intervensie; die relevansie, asook die register en styl, moet in ag geneem word. Aangesien nuwe dokumente, wat in ’n verskeidenheid van genres en vir verskillende kommunikatiewe doeleindes geskryf word, deurentyd op die web gepubliseer of bygewerk word, sal dié sagteware neologismes en opgedateerde terme in ’n bepaalde vakgebied optel. Voorts is hierdie benadering vir terminologiese doeleindes meer effektief en produktief binne vakgebiede wat hoogs gespesialiseerd is, terwyl dit moeiliker is om gespesialiseerde tekste te verkry wat verband hou met populêre vakgebiede, omdat daar ’n oormaat van dié tipe inligting op die web is (Castagnoli 2006:166). Castagnoli 2006:168 meld dat BootCaT wel ’n paar tekortkominge het, o.a. dat die korpussamesteller nie beheer het oor die bronne vanwaar die tekste afkomstig is nie en dat dit nie altyd moontlik is om definisies en bruikbare kontekste vir al die terme te vind nie. Maar die sagteware maak egter daarvoor voorsiening dat die gebruiker sekere filters opstel wat hom/haar toelaat om na elke webtuiste toe te gaan om vas te stel of die inligting gepas is, al dan nie. Binne die Suid-Afrikaanse konteks sal webgebaseerde LSP korpora slegs doeltreffend saamgestel word in Engels en tot ’n mindere mate in Afrikaans, omdat daar ’n tekort is aan (aanlyn-)tekste in die Afrikatale. Ten spyte hiervan weeg die voordele van hierdie eenvoudige en goedkoop benadering tot die versameling van tekste steeds swaarder op teen die tekortkominge.



Figuur 2.2. 'n Skermskoot van die saadwoorde van die vakgebied Maatskaplike Werk.



Figuur 2.3. 'n Skermskoot van die URL's wat onttrek is.

Binne die projek wat die fokus van hierdie studie is, word daar verseker dat die tekste wat in die doelgemaakte korpora opgeneem word, gepaste bronne van primêre data is. 'n Aspek van korpuskompilasie wat nie in bestaande literatuur veel aandag geniet nie, is die gedagte van belyning tussen die aard en inhoud

van die tekste wat versamel word aan die een kant, en die profiel van die teikengebruiker van die uiteindelike terminologiese produk aan die ander kant. Alvorens tekste vir korpuskompilasie versamel word, is dit belangrik om die teikengebruiker te identifiseer, ten einde so ’n belyning moontlik te maak. Ter illustrasie: die tekste wat in ’n korpus opgeneem word wat gebruik gaan word vir die opstel van ’n chemiewoordeboek vir hoërskoolleerders gaan t.o.v. moeilkheidsgraad, teksdigtheid en moontlik selfs register verskil van tekste wat gebruik word as primêre databron vir ’n chemiewoordeboek vir universiteitstudente. Deur voortdurend die teikengebruiker van die terminologiese produk in gedagte te hou, is dit makliker om fundamentele tekste vir die spesifieke gebruiker te identifiseer en in te sluit. So ’n werkswyse lewer ook meer bruikbare gebruiksvoorbeelde op wat ook uit die korpus onttrek word. (Sien hoofstuk 4 vir ’n volledige bespreking oor gebruiksvoorbeelde.) Die teikengebruiker van die uiteindelike terminologiese produk, d.i. die termbank is eerstejaarstudente aan Suid-Afrikaanse universiteite wie se sterkste taal nie Engels is nie. Dié teikengebruiker het nie onderrig in sy/haar sterkste taal ontvang nie, en het ook nie blootstelling aan ’n sterk woordeboekkultuur nie. ’n Verdere voordeel is dat die terminoloog direkte toegang tot vakspesialiste, d.i. dosente het wat ’n oordeel kan vel oor onder andere die toepaslikheid van sekere tekste en aanbevelings in dié verband kan maak.

2.5 Samevatting

Wüster se bydra tot terminologie, wat die GTT-model opgelewer het, het die weggebaan vir die CTT-model. Dié teorie van terminologie blyk superieur te wees oor die GTT wat hoofsaaklik nagestreef het om terme binne bepaalde vakgebiede te standaardiseer. Die CTT maak voorsiening dat vakspesifieke kommunikasie voldoende plaasvind tussen spesialiste en semi-spesialiste – ’n baie belangrike aspek binne die studie. Hoewel toekomstige werk moontlik ’n meer toereikende teorie van terminologie sal oplewer, is die CTT tans die beste kandidaat om die relevante aspekte van ’n term te bestudeer, met behulp van doelgemaakte korpora. Dit is voor die hand liggend dat (doelgemaakte) korpora ’n essensiële rol binne die CTT vervul; binne die Suid-Afrikaanse konteks kan die gebruik van hierdie tipe korpora vir terminologiese doeleindes ’n algemene

praktyk word, mits terminoloë ontvanklik is vir die idee om die hulpbronne wat tot hul beskikking is, te benut.

HOOFSTUK 3

TERMIDENTIFISERING EN SEMI-OUTOMATIESE ONTTREKKING VAN VERKLARENDE INLIGTING

3.1 Inleiding

Vakspesifieke glossariums dien as (a) linguistiese hulpbronne wat die terminologiese basis van ’n vakgebied saamvat, en (b) hulpbronne van kennis waarin definisies van terme weergegee word (Reiplinger et al. 2012:55). In hierdie hoofstuk word die samestelling van sodanige glossariums bespreek deur ’n vergelykende blik te gee op die tradisionele en die moderne benadering tot termidentifisering. Die sagteware vir semi-outomatiese termidentifisering, *WordSmith Tools* en *Sketch Engine*, word met mekaar vergelyk. Laastens word die verskillende benaderings tot die onttrekking van verklarende inligting ook krities bespreek.

3.2 Tradisionele benadering tot termidentifisering

Termidentifisering is die proses waarvolgens relevante terme binne ’n bepaalde vakgebied vanuit ’n gegewe korpus handmatig of semi-outomaties geïdentifiseer word (sien ook paragraaf 1.4). Die tradisionele benadering tot termidentifisering, wat berus op die basiese riglynbeginsel dat konsepte binne ’n bepaalde vakgebied deur terme verteenwoordig word (Alberts 2017:337), behels dat ’n terminoloog potensiële terme handmatig vanuit ’n versameling van (gedrukte) tekste onttrek d.m.v. termbewuste lees. Alvorens ’n terminoloog die terme handmatig kan ekserpeer, moet hy/sy oor voldoende vakkennis beskik, aangesien dié benadering staatmaak op die intuïsie en vakkennis en soms ook algemene kennis van die terminoloog. Dit is nie prakties om dié metode op ’n groot korpus toe te pas nie, omdat dit tydrowend is en verder ook intensiewe konsultasie met ’n vakspesialis behels (Heylyn en De Hertog 2015:199), wat gewoonlik ’n koste-implikasie het. Nchabeleng (2011:5) voeg ook by: “The obvious disadvantage of this method [is that] the margin for human error is rather big.” Dit is wel geskik wanneer terme vanuit ’n kleiner teks onttrek word; ’n duisend woorde word as verstek gebruik (Muegge 2013). Binne die betrokke

benadering word terme nie vanuit elektroniese korpora onttrek nie, maar eerder vanuit handboeke, studiegidse, tydskrifte, ens. (Nchabeleng 2011:5).

3.3 Moderne, korpusgebaseerde benadering tot termidentifisering

Die moderne, korpusgebaseerde benadering tot termidentifisering (hierna verwys as die moderne benadering) berus op die prosessering van elektroniese korpora waaruit potensiële terme onttrek word d.m.v. sagteware vir (semi-) outomatiese termidentifisering. Deur gebruik te maak van korpora en sagteware vir termidentifisering, word terme nie meer in isolasie bestudeer nie, maar kan daar vasgestel word hóé terme in tekste optree. Die konteks waarbinne ’n term gebruik word, voorsien die terminoloog dikwels ook van waardevolle, terminologies-relevante inligting. Rundell (2012:16) voeg by: “... with access to powerful corpus-querying software applied to billion-word corpora, we have the tools (and the data) to provide a fuller and more systematic account of how language works.”

3.3.1 Tegnieke om enkelwoordterme semi-outomaties te identifiseer

Binne die moderne benadering kan twee tegnieke gebruik word om enkelwoordterme binne ’n eentalige korpus (sien parargraaf 1.4) te identifiseer, nl.: (a) deur al die woorde in die korpus te tel, sal die sagteware ’n lys oplewer wat potensiële terme bevat wat aan die minimum frekwensiewaarde voldoen wat deur die terminoloog gestel is; en (b) om ’n doelgemaakte korpus met ’n algemene verwysingskorpus te vergelyk (Bowker en Pearson 2002:166-167). Eersgenoemde proses lewer bloot ’n lys van alle woorde wat in die korpus voorkom op. Gewoonlik kan so ’n lys volgens frekwensie gesorteer word. Die terminoloog kan besluit wat die minimum frekwensie is vir woorde om in die lys ingesluit te word – Bowker en Pearson (2001:166) dui die verstekwaarde as vyf keer aan. Hierdie minimum vereiste kan aangepas word na gelang van die grootte van die korpus. Wanneer die frekwensiewaarde te hoog gestel word sal terme wat dikwels verskyn, maar in verskillende vorme, moontlik nie in die termlys gelys word nie (Bowker en Pearson 2002:167).

As strategie vir termonttrekking het hierdie proses beperkte gebruikswaarde. Dit is alombekend dat sogenaamde funksiewoorde die hoogste frekwensie binne

enige korpus het. Funksiewoorde soos lidwoorde, voorsetsels en voegwoorde is baie selde terme, maar sal tog bo-aan so ’n frekwensielys verskyn. Vergelyk in dié verband ’n woordelys gebaseer op ’n doelgemaakte korpus van die projek vir Kommunikasiepatologie (figuur 3.1):

1	the	107141
2	of	68399
3	a	48468
4	and	43538
5	in	40935
6	to	39157
7	is	38383
8	that	21371
9	are	16661
10	as	15982
11	be	15578
12	The	14877
13	for	14710
14	or	11890
15	with	11885
16	it	10550
17	not	9739
18	by	9579
19	o	9252
20	can	8682
21	this	8241
22	an	7880
23	e	7822
24	on	7693
25	which	7345
26	have	7247
27	f	7118
28	speech	6832
29	we	6823
30	from	6391

Figuur 3.1. ’n Skermskoot van die woordelys van Kommunikasiepatologie.

’n Verdere beperking is dat so ’n woordelys met ongelemmatiseerde items werk. Ter illustrasie: hoewel ’n menslike leser kan herken dat byvoorbeeld “kontantvloeistaat”, “Kontantvloeistaat” en “kontantvloeistate” na dieselfde konsep verwys en derhalwe een lemma verteenwoordig, sal die meeste sagteware elke leksikale item afsonderlik tel. So kan dit gebeur dat verskillende leksikale items wat dieselfde lemma verteenwoordig as individuele items nie die minimum frekwensievereiste haal nie, en daarom nie op so ’n frekwensielys verskyn nie.

Die tweede tegniek, dit wil sê die vergelyking van ’n doelgemaakte korpus met ’n algemene verwysingskorpus is veel meer effektief, aangesien dié tegniek op ’n statistiese benadering berus (Baker 2004:346). Die *KeyWord*-funksie van

byvoorbeeld *WordSmith Tools*, wat sleutelwoorde identifiseer, werk op hierdie beginsel (Taljard en De Schryver 2002:51, McEnery et al. 2006:308).

Scott (1997:26) definieer ’n sleutelwoord as “a word which occurs with unusual frequency in a given text.” Dit impliseer dus dat woorde wat nie noodwendig ’n groot aantal kere met ’n hoë, werklike frekwensie in die doelgemaakte korpus voorkom nie, wel ’n hoë relatiewe frekwensie sal hê wanneer dit met ’n algemene korpus vergelyk word. In hierdie geval sal die algemene verwysingskorpus altyd groter as die doelgemaakte korpus wees. Die identifisering van sleutelwoorde met behulp van die *KeyWord*-funksie binne *WordSmith Tools* word stapsgewys uitgevoer. Scott (1997:236) en Taljard en De Schryver (2002:51-52) meld dat daar eerstens ’n woordelys van die algemene verwysingskorpus saamgestel word deur middel van die *WordList*-funksie. So ’n woordelys bevat al die verskillende tipes (Eng: *types*) in die algemene verwysingskorpus en word volgens (die hoogste) frekwensie gerangskik. Tweedens word ’n soortgelyke woordelys van die doelgemaakte korpus saamgestel, wat heelwat kleiner as die woordelys van die verwysingskorpus. Taljard en De Schryver (2002:52) dui aan dat die tweede woordelys saamgestel word uit “specific text(s) for which the key words are to be identified.” Laastens word die *KeyWord*-funksie gebruik om die frekwensie van elke item in die kleinste woordelys met die frekwensie van elke item in die woordelys van die algemene verwysingskorpus, te vergelyk. Scott (1997:236) meld dat items wat ’n groot verskil in frekwensie het, as sleutelwoorde geïdentifiseer word; items met dieselfde frekwensie mag moontlik nie geïdentifiseer word nie. (Vir ’n volledige bespreking van die presiese statistiese prosedure van die *KeyWord*-funksie van *WordSmith Tools*, sien Taljard en De Schryver (2002: 52 et seq.)).

Sketch Engine gebruik ook statistiese bewerkings vir termonttrekking (https://www.sketchengine.co.uk/wp-content/uploads/ske-statistics.pdf).
Kilgarriff et al. (2014:30) noem dat die statistiese metode wat gebruik word om sleutelwoorde te identifiseer ook geskik is om terme te identifiseer.

Ten einde die verskil tussen ’n blote frekwensielys en ’n *KeyWord*-lys te illustreer, word die *KeyWord*-lys van die doelgemaakte korpus vir Kommunikasiepatologie weergegee (figuur 3.2):

	NUMBER	ITEM	SCORE	FREQ	REF_FREQ
2	1	vowel	689.13	3325	12382
3	2	vowels	553.15	2505	10831
4	3	velopharyngeal	492	1219	66
5	4	cleft	419.29	2186	14435
6	5	formant	380.48	988	699
7	6	phonetic	357.71	1539	9656
8	7	consonants	338.26	1247	6425
9	8	phonetics	333.54	990	2654
10	9	velum	299.85	756	312
11	10	articulation	276.88	1799	21191
12	11	articulatory	264.04	677	542
13	12	pharyngeal	259.97	736	1945
14	13	fricatives	244.5	620	398
15	14	consonant	242.84	974	8143
16	15	voiceless	211	686	4162
17	16	fricative	207.56	535	627
18	17	syllable	204.89	999	12694
19	18	utterance	198.36	907	11104
20	19	alveolar	188.11	600	3847
21	20	glottal	183.92	478	747

Figuur 3.2. ’n Skermskoot van die *KeyWord*-lys van Kommunikasiepatologie.

Wanneer figuur 3.1 en figuur 3.2 met mekaar vergelyk word, is die verskil ooglopend. In die *WordList* (figuur 3.1) word die eerste 27 posisies deur funksiewoorde beklee; eers in posisie 28 word die eerste item, *speech,* met leksikale betekenis aangetref. Die enigste manier waarop terme en funksiewoorde in ’n frekwensielys van mekaar onderskei kan word, is handmatige onttrekking deur ’n terminoloog en / of ’n vakspesialis. Dit illustreer dus duidelik die beperkte gebruikswaarde van blote frekwensielyste vir terminologiese doeleindes. In figuur 3.2 daarteenoor is die termstatus van al die items op die lys, beginnende by posisie 1 reeds duidelik. Al hierdie items verwys na konsepte binne die veld van Kommunikasiepatologie. Hierdie konsepte kan dus in die veeltalige aanlyn termbank vir die teikengebruiker aangebied word.

3.3.2 Tegnieke om meerwoordterme semi-outomaties te identifiseer

Twee benaderinge kan gevolg word om meerwoordterme te identifiseer. Bowker en Pearson (2002:168) verwys in dié verband na die linguistiese en statistiese metodes. Elkeen van hierdie benaderinge word vervolgens bespreek.

Die linguistiese benadering is gebaseer op die beginsel dat meerwoordterme die neiging het om spesifieke (morfo-)sintaktiese patrone te vertoon. “They (multiword terms) rely on this templatic behaviour to determine the validity of a word combination as a linguistic unit, and if so, as a TC (term candidate)” (Heylyn en De Hertog 2015:205). In Engels byvoorbeeld, is naamwoord + naamwoord en

adjektief + naamwoord algemene sintaktiese patrone vir meerwoordterme, byvoorbeeld *business cheque* en *capital account*. Binne hierdie benadering moet elke woord in die korpus geëtiketteer word met die toepaslike woordsoort – ’n geannoteerde korpus dus. Wanneer die woorde geëtiketteer is, sal die sagteware ’n termlys onttrek wat potensiële terme bevat wat met die gespesifiseerde woordsoortpatrone ooreenstem. Die nadeel van dié benadering is dat daar ook geraas en stilte (sien paragraaf 1.4) in termlyste mag voorkom, asook ellipses (wanneer ’n woord geïmpliseer word, maar dit word nie eksplisiet gemeld nie). ’n Verdere ooglopende nadeel van hierdie metode is dat dit slegs op geannoteerde tekste toegepas kan word. Vir tale wat arm is aan hulpbronne in mensliketaaltegnologie, soos die Afrikatale, is hierdie metode dus van weinig waarde.

Aan die ander kant maak die statistiese benadering staat op die sagteware om na iteratiewe reekse van leksikale items te soek (Bowker en Pearson 2002:170). ’n Minimum frekwensiewaarde word deur die terminoloog gestel wat die aantal kere bepaal wat ’n reeks items herhaal moet word, voordat dit as ’n potensiële term in die termlys gelys word. Binne dié benadering sal daar ook geraas en stilte in die termlyste voorkom; stoplyste (’n lys wat woorde bevat wat uitgesluit moet word wanneer terme geïdentifiseer word) kan wel gebruik word om die aantal ongeldige terme in die termlys te verminder. Die voordeel verbonde aan dié benadering is dat dit nie staatmaak op linguistiese inligting nie en nie taalspesifiek is nie: “… a term extraction tool employing this approach can, in principle, be used to process texts in multiple languages” (Bowker en Pearson 2002:171).

Die persepsie bestaan onder terminoloë dat die moderne benadering sekere leksikale items soos afkortings, sinonieme en neologismes, nié sal onttrek nie en dat dit slegs handmatig geëkserpeer kan word. Maar, soos Alberts (2017:337) meld, is die teendeel waar: “Machine processing also picks up lexical items which have a relatively low frequency of occurrence but which might be of particular importance within a specific language or subject field.”

Die prosedure vir semi-outomatiese termonttrekking soos hierbo beskryf word vervolgens aan die hand van ’n doelgemaakte korpus vir rekeningkunde geïllustreer.

3.4 Semi-outomatiese termonttrekking: ’n gevallestudie

In die betrokke gevallestudie word leksikale items wat beskou word as items met ’n lae frekwensie, geselekteer. Vanweë die lae frekwensie van sulke items, word daar dikwels aanvaar dat hulle slegs handmatig geëkserpeer kan word. Daar word vervolgens aangedui dat dit wel moontlik is om items met ’n lae frekwensie semi-outomaties te onttrek.

’n Lys van rekeningkundeterme wat binne die projek gebruik is, dien as bewys van goeie belyning tussen ’n korpus (die aard en inhoud van die tekste) en die profiel van die veronderstelde teikengebruiker van die terminologiese eindproduk (dit sluit in die vaardighede, vakkennis en taalkennis van die teikengebruiker), waar die terme met behulp van sagteware semi-outomaties onttrek is. Die korpus is saamgestel met eerstejaaruniversiteitstudente as teikengebruikers van die finale produk, d.i. die termlys met sy definisies (sien paragraaf 1.1. i.v.m. die akademiese gereedheid van ’n matriekleerling vir universiteit). Die tekste wat in die korpus opgeneem is, is deur vakspesialiste, verbonde aan die Universiteit van Pretoria, verskaf. Die tekstipes, wat nie meer as ’n jaar oud is nie, is studiemateriaal en word in ’n eerstejaarrekeningkundekursus gebruik. Die aanvanklike grootte van die korpus is 31 193 woorde, maar die korpus is organies en kan uitgebrei word. Die rekenaarmatige proses van termidentifisering is uitgevoer met behulp van die sagteware *Sketch Engine*. As algemene verwysingskorpus, is die *English Web 2013 (enTenTen13)* gebruik. Die betrokke algemene verwysingskorpus is in die sagteware ingebou en spaar dus die terminoloog die moeite om self ’n verwysingskorpus saam te stel. Die terme in die termlys is gerangskik vanaf die hoogste na die laagste frekwensiewaarde. Dit is belangrik om te meld dat die resultaat nie net ’n blote frekwensielys is nie, maar eerder ’n kandidaattermlys wat die resultaat van ’n statistiese bewerking is. Die kandidaattermlyste is daarna deur twee vakspesialiste gevalideer wat die geraas verwyder het; die ‘stiltes’ is sover moontlik opgevul deur die voltooiing van bepaalde konseptuele paradigmas - ’n

ingreep wat slegs deur vakspesialiste uitgevoer kan word. Die gevalideerde enkelwoordtermlys bevat altesaam 117 terme waarvan 13 terme deur die vakspesialiste bygevoeg is. Voorbeelde van geraas wat in die betrokke termlys voorgekom het sluit in eiename, omdat voorbeelde en oefeninge waarin daar na denkbeeldige persone en instansies verwys word, in die studiemateriaal gebruik word.

	NUMBER	ITEM	SCORE	FREQ	REF_FREQ
1	1	account	44.85	326	2204422
2	2	VAT	797.95	245	80692
3	3	accounting	157.42	209	392047
4	4	credit	14.67	199	4126056
5	5	cash	30.35	193	1926753
6	6	income	31.16	165	1602386
7	7	statement	48.23	165	1030700
8	8	profit	71.12	154	647616
9	9	amount	14.06	131	2829105
10	10	journal	65.77	121	548312
11	11	cost	11.25	116	3133461
12	12	assets	54.19	110	606426
19	18	SARS	757.41	62	12015
20	19	creditor	88.61	56	179935

Figuur 3.3 ’n Skermskoot van die enkelwoordtermlys van rekeningkunde.

Die gevalideerde meerwoordtermlys bevat altesaam 315 terme; 32 terme is deur die vakspesialiste bygevoeg. Ongeldige terme wat geïdentifiseer is sluit terme in waarin ’n woord herhaal word of die ontbreking van een of meer woorde in ’n bepaalde meerwoordterm. Daar is bevind dat die leksikale items wat beskou word as items wat heel waarskynlik ’n lae frekwensie het en nie rekenaarmatig onttrek sal word nie, wél binne die betrokke gevallestudie onttrek is, byvoorbeeld die akroniem *VAT* (sien figuur 3.3) en sinonieme in die meerwoordtermlys (sien figuur 3.4). Trouens, eersgenoemde leksikale item het die tweede hoogste frekwensie in die enkelwoordtermlys en dus kan die aanname nie meer gemaak word dat sekere leksikale items ’n lae frekwensie het en nie semi-outomaties onttrek kan word nie.

	NUMBER	ITEM	SCORE	FREQ	REF_FREQ
24	22	petty cash	249.14	15	5450
25	23	reconciliation statement	326.34	14	162
26	24	duality concept	306.43	13	14
27	25	accounting system	188.69	12	6507
28	26	interest income	183.31	12	7072
29	27	inventory account	281.81	12	76
30	28	post-adjustment trial balance	283.22	12	0
31	29	bank reconciliation statement	257.9	11	92
32	30	double-entry system (synonym for lines 116, 117, 118, 119)	257.9	11	91
33	31	final consumer	249.71	11	520
34	32	accumulated depreciation	235.01	10	71
35	33	perpetual inventory system	235.71	10	38
36	34	capital account	186.88	9	1793
37	35	debtors control	212.66	9	0
38	36	financial performance	121.94	9	9651
39	37	conceptual framework	140.21	8	4532
40	38	interest expense	142.64	8	4233
41	39	long-term loan	177.6	8	852
42	40	credit entry	163.18	7	203
43	41	credit policy	149.62	7	1399
44	42	income receivable	165.3	7	35
45	43	periodic inventory system	165.46	7	16
46	44	post-closing trial balance	165.63	7	6
47	45	double-entry principle (synonym for lines 116, 117, 118, 119)	142.11	6	3
48	46	financial accounting	108.32	6	4057

Figuur 3.4 ’n Skermskoot van die meerwoordtermlys van rekeningkunde.

Dit is moeilik om ’n statisties goed gefundeerde evaluering van die semi-outomatiese proses van termonttrekking te maak, aangesien daar geen handmatige ekserperingsresultate is waarmee die resultate van ’n semi-outomatiese proses vergelyk kan word nie - vergelyk in dié verband die analise van *recall* en *precision* deur Taljard en De Schryver (2002:55). Tog gee ’n eenvoudige analise van die resultaat van semi-outomatiese termonttrekking ’n aanduiding van die relatiewe effektiwiteit daarvan. Van die potensiële 113 enkelwoordterme is 75% wel suksesvol uit die korpus onttrek, en dit moet verder in gedagte gehou word dat die terme wat deur vakkundiges bygevoeg is, nie noodwendig in die korpus voorkom nie – van hulle is bygevoeg om bepaalde konseptuele paradigmas te voltooi. Vir die meerwoordterme is die resultate selfs beter – 90% van potensiële terme is semi-outomaties onttrek. Wat wel duidelik is, is dat die omvang van menslike intervensie aansienlik minder is as dié van die tradisionele benadering, waar daar swaar gesteun word op die insette van vakspesialiste.

3.5 *WordSmith Tools* en *Sketch Engine*: ’n vergelyking

Sagteware vir termidentifisering word gebruik om vas te stel hóé woorde in ’n versameling van tekste gebruik word. Drie soorte sagteware vir termidentifisering

is beskikbaar (Costa et al. 2016:15-17): losstaande sagteware, webgebaseerde sagteware en herbruikbare sagteware, wat bekendstaan as raamwerke (Eng: *frameworks*). Losstaande sagteware, soos byvoorbeeld *Simple Extractor* (www.dail-software.com), word op ’n rekenaar geïnstalleer en werk onafhanklik van enige ander sisteem, byvoorbeeld ’n internetkonneksie, of enige ander elektroniese toestel. Hierdie tipe sagteware bied gewoonlik individuele of kommersiële lisensies aan wat eenmalig aangeskaf word - wat op die lang duur die goedkoopste opsie is - en dit kan vir ’n onbeperkte tydperk gebruik word. Die gebruiker sal altyd toegang tot sy/haar data hê en is self verantwoordelik vir rugsteunkopieë van byvoorbeeld die verwysingskorpora. Web-gebaseerde sagteware hoef nie op ’n rekenaar geïnstalleer te word nie en toegang tot hierdie sagteware word verkry d.m.v. ’n webblaaier. Die gebruiker kan enige tyd, enige plek toegang tot die sagteware verkry; toegang mag egter belemmer word gedurende staantyd (Eng: *downtime*). Individuele of kommersiële subskripsies is beskikbaar en as ’n reël moet dié subskripsies jaarliks hernu word. Die koste-implikasies moet daarom in ag geneem word wanneer ’n keuse uitgeoefen word oor watter sagteware om aan te skaf. Gepaardgaande met die subskripsie, het die gebruiker toegang tot ’n reeks bygewerkte algemene verwysingskorpora – ’n funksie wat by losstaande sagteware ontbreek en wat self deur die gebruiker verskaf moet word wanneer byvoorbeeld sleutelwoorde onttrek word. ’n Voorbeeld van hierdie tipe sagteware is *Terminus* (terminus.iula.upf.adu). Raamwerke, soos byvoorbeeld Okapi Framework (okapi.sourceforge.net), is nie volledige sagteware nie, maar wel herbruikbare sagteware wat geïntegreer kan word in ander tipe sagteware. “... systems of this type are often used in information retrieval, where identification and indexing of terminology serves as an aid to information retrieval queries” (Costa et al. 2016:16).

Om vas te stel of losstaande- en web-gebaseerde sagteware dieselfde resultate oplewer tydens die termidentifiseringsproses, is *WordSmith Tools* en *Sketch Engine* as voorbeelde van losstaande- en webgebaseerde sagteware onderskeidelik met mekaar vergelyk. Vir die doel van die eksperiment is ’n rekeningkundekorpus gebruik. Om die enkelwoordterme te onttrek, is die *University of Pretoria’s English Corpus*, wat uit 290 000 woorde bestaan, binne *WordSmith Tools* as algemene verwysingskorpus gebruik; binne *Sketch Engine*

het die *English Web 2013 (enTenTen13)* gedien as algemene verwysingskorpus. ’n Honderd enkelwoordterme en 100 meerwoordterme is semi-outomaties met behulp van *WordSmith Tools* en *Sketch Engine* onttrek. Die minimum frekwensiewaarde is op vyf gestel en die termlyste is volgens frekwensie gesorteer. Vanuit die twee enkelwoordtermlyste stem 51 uit 100 terme met mekaar ooreen (sien figuur 3.6; ooreenstemmende terme is in groen gemerk).

NUMBER	SKETCH ENGINE	SCORE	FREQ	REF_FREQ		WORDSMITH	Freq.	RC. Freq.
1	vat	797.95	245	80692	1	ENTITY	329	91
2	sars	757.41	62	12015	2	FINANCIAL	295	186
3	ledger	727.07	104	30676	3	VAT	234	12
4	fol	673.37	44	6980	4	ACCOUNTING	204	28
5	debtors	627.17	125	47840	5	R	237	440
6	entity	431.81	343	229326	6	ACCOUNT	318	2094
7	gj	421.86	22	2968	7	CASH	193	208
8	ledgers	405.87	25	5850	8	SALES	161	118
9	debited	334.5	20	5306	9	CREDIT	195	525
10	buba	298.57	13	355	10	INVENTORY	130	13
11	crj	284.53	13	1012	11	BANK	229	1246
12	liabilities	264.62	76	74675	12	DEBTORS	128	16
13	diagram	236.22	83	94251	14	BALANCE	174	334

Figuur 3.5. ’n Skermskoot van ‘n gedeelte van die ooreenstemmende terme wat semi-outomaties onttrek is.

Een moontlike rede waarom daar nie groter ooreenstemming tussen die twee kandidaattermlyste is nie, is omdat twee verskillende verwysingskorpora, wat nie dieselfde aantal woorde bevat nie, onderskeidelik gebruik is. Dit was opvallend dat uit die 51 terme sowel enkelvoud- as meervoudsvorme van sekere terme onttrek is. Voorbeelde hiervan sluit in: *debtor* en *debtors*, *ledger* en *ledgers*, *transaction* en *transactions*. Dit bevestig dat beide *WordSmith Tools* en *Sketch Engine* nie dieselfde konsepte saamgroepeer nie en dat menslike intervensie, soos wat dit deurentyd in die onderhawige studie beklemtoon word, nodig is (hetsy met die programmering van die sagteware of die verwydering van oortollige terme). Lemmatisering moet ook handmatig gedoen word. Daar is ook gekyk of die losstaande sagteware enige van die 13 terme onttrek het wat deur die vakspesialiste bygevoeg is. Slegs 2 van die 13 terme is in die betrokke enkelwoordtermlys gevind. Die ander 11 terme het moontlik nie aan die minimum frekwensiewaarde voldoen nie en is dus nie rekenaarmatig onttrek nie. Die meerwoordtermlyste het geen resultate opgelewer nie. Hoewel vroeër

weergawes van *WordSmith Tools* meerwoordterme kon onttrek, blyk dit dat dit nie moontlik is om dié funksie binne nuwer weergawes van die betrokke sagteware uit te voer nie (binne die onderhawige studie is weergawe 6.0 gebruik). Daar kon dus nie ’n vergelyking tussen die data van die meerwoordterme vanuit die onderskeie sagteware getref word nie.

In die resultate van die vergelyking van die losstaande- en webgebaseerde sagteware, lewer beide sagteware terminologies relevante data op. Dit is egter nie moontlik om ’n betroubare, kwantitatiewe vergelyking tussen die twee tipes sagteware te maak nie, omdat daar nie ’n handmatig geëkserpeerde termlys is waarteen *recall* en *precision* gemeet kan word nie.

3.6 Saamstel van definisies

’n Volgende logiese stap in die opstel van ’n terminologiese databank is die byvoeg van definisies van die geverifieerde terme wat semi-outomaties geïdentifiseer is.

Temmerman (2000:9) noem dat ’n terminologiese definisie die plek van ’n konsep binne ’n sisteem van konsepte aandui; terminologiese definisies onderskei dus tussen verwante konsepte binne ’n konseptuele sisteem. Alberts (2017:73) meld dat die konsep, sy definisie en term aan dieselfde vakgebied moet behoort: “The term-definition equation is inadequate by itself and needs a relationship with a subject field in order to provide the link to extra-linguistic activity. It is this association with a specific subject field which confirms that a concept and term belong to the same conceptual cluster or system.”

Met die profiel van die veronderstelde teikengebruiker in ag genome is dit belangrik om binne ’n projek wat daarop gemik is om universiteitstudente kognitief te ondersteun, terme met gepaardgaande definisies te verskaf. Reiplinger et al. (2012:55) voeg by dat definisies die begrip van terme bevorder.

Dit is belangrik om in gedagte te hou dat terminologiese definisies en verklarende inligting nie dieselfde is nie. Vir die doel van hierdie studie word verklarende inligting gedefinieer as enige inligting wat rekenaarmatig binne ’n doelgemaakte korpus gevind kan word wat verband hou met die betekenis en die gebruik van sowel ’n term, as ’n term se konseptuele verhoudinge met ander terme (Taljard

2004:176). Hierdie inligting kan deur die terminoloog gebruik word om ’n terminologiese definisie te formuleer, wat dan aan ’n vakspesialis voorgelê kan word vir kontrole en validering.

Daar bestaan verskillende strategieë vir die formulering van terminologiese definisies. In die eerste plek kan definisies geformuleer word met behulp van ’n vakspesialis. In die tweede plek kan definisies opgesoek word in ’n verskeidenheid vakspesifieke woordeboeke, waarna die verskillende definisies van ’n bepaalde term gekombineer word om ’n nuwe definisie te vorm. In die derde plek kan definisies geformuleer word met verklarende inligting wat semi-outomaties uit ’n doelgemaakte korpus onttrek word as basis / vertrekpunt. Elkeen van hierdie strategieë word vervolgens bespreek.

3.6.1 Samewerking met vakspesialiste

Wanneer ’n terminoloog definisies formuleer met behulp van ’n vakspesialis, behels dit intensiewe konsultasie met sodanige vakspesialis wat potensieel duur en tydrowend is. Die ideaal sou wees dat verskeie vakspesialiste betrokke is, sodat definisies nie ’n eensydige, subjektiewe siening reflekteer nie, wat tot voordeel van die teikengebruiker sal wees. “Different experts may give conflicting advice on concepts and terms … LSP learners would benefit from consulting multiple experts in a given subject field, though in practice, this is not always a realistic aim” (Bowker en Pearson 2002:17-18). Hoewel die verwysing na “LSP learners” is, geld die advies ook vir terminoloë. Taljard (2004:177) wys daarop dat ’n terminoloog oor ’n baie spesifieke vaardigheid moet beskik om die korrekte en toepaslike inligting van ’n vakspesialis te verkry. Sy voeg by dat ’n terminoloog nie noodwendig oor die vakkennis beskik om tussen relevante en irrelevante inligting te onderskei nie en gevolglik nie ’n oordeel kan vel oor moontlike subjektiewe inligting nie. Die terminoloog behoort ideaal gesproke oor voldoende vakkennis binne ’n bepaalde vakgebied te beskik om so ver as moontlik akkurate data in te samel. Dit is ook belangrik dat terminoloë opgelei moet word in hóé om onderhoude met vakspesialiste te voer, veral binne die Suid-Afrikaanse konteks, omdat terminologie-opleiding nog in sy kinderskoene staan (Taljard 2004:177).

3.6.2 Hergebruik, uitbreiding en herbewerking van bestaande definisies

Om vakspesifieke woordeboeke te raadpleeg om verskillende definisies vir terme te versamel en te kombineer om nuwe definisies te vorm, is erkende leksikografiese praktyk. Kopiereg en die beginsel van intellektuele eiendom moet egter deurgaans deur die terminoloog in ag neem word. In die geval van terminologiese definisies moet dit egter in gedagte gehou word dat terme teoreties gesproke eenduidig is en dat ’n bepaalde definisie net op ’n beperkte aantal maniere geformuleer kan word: dit sou byvoorbeeld moeilik wees om die term ‘reguitlyn’ op enige ander wyse te formuleer as ‘die kortste afstand tussen twee punte’. Dit is ook belangrik dat die terminoloog veelvuldige, gesaghebbende bronne raadpleeg en nie bronne lukraak kies nie. Voorts staar die terminoloog ’n paar uitdagings in die gesig. Bowker en Pearson (2002:15) noem dat een van die grootste tekortkominge van vakspesifieke woordeboeke toegeskryf word aan hulle onvolledigheid. Hulle voeg by dat met die snelle ontwikkeling van vakgebiede soos tegnologie en wetenskap, die moontlikheid bestaan dat wanneer ’n vakspesifieke woordeboek geraadpleeg word, dit nie die huidige stand van sake binne ’n bepaalde vakgebied sal reflekteer nie. Dit geld veral vir papierwoordeboeke. ’n Moontlike rede waarom die inhoud van gedrukte woordeboeke as gedateerd beskou word, is die feit dat dit ’n langdurige proses is om woordeboeke saam te stel én te publiseer. Met die ontwikkeling van aanlyn woordeboeke sal hierdie probleem uiteraard minder ernstig raak, aangesien bywerking van elektroniese vakwoordeboeke – in teorie, in elk geval – onmiddellik gedoen kan word. Nog ’n beperking binne vakspesifieke woordeboeke is dat kontekstuele- en gebruiksinligting dikwels nie weergegee word nie – inligting wat noodsaaklik is vir die teikengebruiker (Bowker en Pearson 2002:16). Dit is ook belangrik dat die terminoloog tydens die skryf van definisies die veronderstelde teikengebruiker deurentyd in gedagte hou om te verseker dat daar ’n belyning tussen die inhoud en moeilikheidsgraad van die definisies en die teikengebruiker is: “… there is always the danger of incompatibility of the target group served by the dictionary and the target group which the terminologist has in mind” (Taljard 2004:177). Indien die terminoloog daarin slaag om definisies te versamel wat verband hou met die profiel van die veronderstelde teikengebruiker, kan hy/sy begin deur die relevante definisies met mekaar te

kombineer om nuwe definisies te skep. Die risiko hieraan verbonde is dat die terminoloog inligting in- of uitsluit op grond van sy/haar intuïsie en nie op grond van vakkennis en die behoeftes van die veronderstelde teikengebruiker nie; dit kan dus lei tot onvolledigheid of selfs foutiewe ensiklopediese inligting. Hoewel hierdie benadering die insette van die vakspesialis verminder, bly die konsultasie met spesialiste onontbeerlik.

3.6.3 Semi-outomatiese onttrekking van verklarende inligting

Pearson (1998:135) is van mening dat wanneer ’n vakspesialis vakspesifieke tekste saamstel, hulle (bewustelik/onbewustelik) sekere terme sal verduidelik: “The extent to which they do this will depend on the perceived disparity of knowledge between the author and reader.” Sy voeg by dat dit gedeeltelike of volledige verduidelikings mag wees en dat dié inligting (semi-outomaties) onttrek kan word vir die samestelling van vakspesifieke definisies. Binne die benadering van semi-outomatiese onttrekking van verklarende inligting, raadpleeg die terminoloog ’n doelgemaakte korpus waaruit hulle heel waarskynlik die mees onlangse inligting oor ’n bepaalde term sal kan onttrek. Taljard (2004:178) meld dat so ’n korpus oorvloedige tekstuele- en gebruiksinligting bevat: “Should the terminologist wish to include examples of use as part of the definition of a term, authentic examples can be sourced directly from the corpus.”

Reiplinger et al. (2012:56) noem dat die meeste tegnieke vir semi-outomatiese identifisering van verklarende inligting gebaseer is op die identifisering van naamwoorde en/of leksikale patrone wat tipies in definisies voorkom. Hulle voeg by dat die patrone rekenaarmatig gefiltreer kan word op grond van die kenmerke wat verwys na die sintaktiese konteks, leksikale inhoud, leestekens, byvoorbeeld hakies en koppeltekens, uitleg, posisie in diskoers, ens.

Twee tegnieke kan gebruik word om verklarende inligting semi-outomaties te onttrek, nl.: (a) die algoritme-gebaseerde metode (Pearson 1998:135) en (b) die *KeyWord*-in-Context-gebaseerde metode (afgekort as KWIC-gebaseerde metode) (Taljard 2004:180). Die twee tegnieke word vervolgens bespreek.

Pearson (1998:135) tref ’n onderskeid tussen twee tipes verklarende inligting in vakspesifieke tekste, d.i. inligting waar die term en sy definisie in dieselfde sin

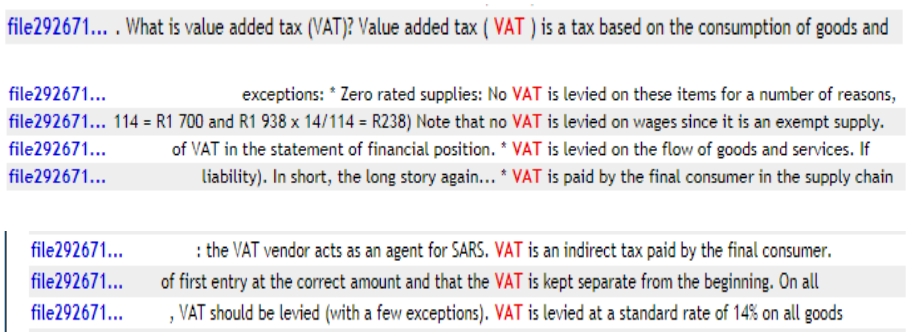
verskyn, en inligting waar die term en sy definisie oor twee sinne versprei is. Sy gebruik die terme ‘formele verklarende inligting’ en ‘informele verklarende inligting’ onderskeidelik in dié verband. Sy vervolg dat verklarende inligting ’n formulêre karakter het, en dat die herwinning van verklarende inligting op die rekenaarmatige herkenning van hierdie formule in ’n teks berus. Pearson (1998:136 et seq.) identifiseer twee formules deur middel waarvan ’n formele definisie uitgedruk kan word:

- Formule 1: *X* = *Y* + kenmerkende eienskap, waar *X* die subordinaat/ondergeskikte van *Y* is. Vergelyk: ’n bywoord (*X*) is (=) ’n linguistiese woord (*Y*) wat ’n werkwoord omskryf (kenmerkende eienskap). Die bywoord (*X*) is dus die subordinaat van die superordinaat ‘linguistiese woord’ (*Y*), en die kenmerkende eienskap wat bywoorde van ander linguistiese woorde onderskei, is die feit dat hulle werkwoorde omskryf (Taljard 2004:182).
- Formule 2: *Y* + kenmerkende eienskap = *X*, waarby *X* die subordinaat van *Y* is, byvoorbeeld (Taljard 2004:182): ’n linguistiese woord (*Y*) wat ’n werkwoord omskryf (kenmerkende eienskap) is (=) ’n bywoord (*X*). Die bywoord (*X*) is dus die subordinaat van die superordinaat ‘linguistiese woord’ (*Y*), en die kenmerkende eienskap wat bywoorde van ander linguistiese woorde onderskei, is die feit dat hulle werkwoorde omskryf (Taljard 2004:182).

In die bostaande formules verteenwoordig *X* die geïdentifiseerde term (’n term wat deur die vakspesialis geverifieer is), *Y* die superordinaat en = ’n linguistiese eenheid. Pearson spesifiseer ook vullers vir die gleuwe *Y* en =. Die gleuf *Y* moet óf ’n term óf ’n reeks gespesifiseerde woordklasse wees (Pearson 1998:138). Indien *Y* ’n woordklas is, verwys dit na tipiese terme, soos: *funksie, tegniek, metode, proses, klas*. Die gleuf = mag gevul word deur wat Pearson (1998:139) ’n ‘connective’ noem: “We have chosen to use the general term *connective* to describe all verbs and phrases which connect a term with information about the term; such information may range from the provision of a formal defining expositive to simply the specification of the term’s superordinate.” Tipiese konnektiewe wat in hierdie gleuf gevind word, is: *is, word genoem, bestaan uit,*

gedefinieer as, staan bekend as. Dié woorde dui ’n moontlike kenmerkende eienskap of eienskappe aan van dit wat volg. Die algoritme-gebaseerde metode maak egter nie voorsiening daarvoor dat verklarende inligting vanuit tale wat hulpbron-arm is, onttrek word nie; dié benadering kan dus slegs toegepas word in ’n Engelse korpus.

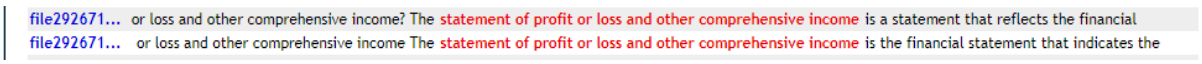
Die KWIC-gebaseerde metode berus op ’n lys van konkordansielyne of sleutelwoorde-in-konteks-lyne (in Eng: *KeyWords-in-context*, afgekort as KWIC), wat verkry word d.m.v. sagteware se konkordansiefunksie. Die KWIC-lyne sal outomaties verskyn wanneer daar na ’n bepaalde term binne die korpus gesoek word; dus word die bepaalde term binne die tekstuele konteks geplaas. Deur die KWIC-lyne te bestudeer (wat links of regs gesorteer kan word of sekere filters wat gebruik kan word om inligting in- of uit te sluit), kan daar vasgestel word of verklarende inligting van ’n bepaalde term binne die korpus weergegee word (Taljard 2004:180). Daar moet in gedagte gehou word dat die aantal KWIC-lyne van een term na die volgende sal verskil, afhangende van die frekwensie van ’n bepaalde term binne ’n korpus. Ter illustrasie is twee terme lukraak uit die geverifieerde rekeningkundelys gekies – een term wat rekenaarmatig onttrek is, *VAT* en een term, *statement of profit or loss and other comprehensive income*. Die konkordansiefunksie van die sagteware *Sketch Engine* is gebruik om na die betrokke terme binne die doelgemaakte korpus te soek. Die KWIC-lyne is regs gesorteer, dus verskyn die resultate alfabeties en vergemaklik dit die bestudering daarvan. Vergelyk figuur 3.3 vir ’n greep van die 245 KWIC-lyne wat vir die term *VAT* opgelewer is. In 8 van die lyne is verklarende inligting gevind.Vergelyk figuur 3.3 hieronder vir die KWIC-lyne vir die term *VAT* (vir die doel van die bespreking is die KWIC-lyne wat verklarende inligting bevat, geïsoleer):



Figuur 3.3. Geïsoleerde KWIC-lyne van die term *VAT.*

In die eerste KWIC-lyn is ’n prototipiese voorbeeld van die eerste formule, soos hierbo uiteengesit. Die term *VAT* wat as soekwoord gebruik is, verteenwoordig *X,* die superordinaat *Y* is *tax*, en die konnektiewe is *based on the consumption of goods and services*. Ander inligting wat potensieel bruikbaar is vir die saamstel van ’n definisie is *an indirect tax paid by the final consumer* (KWIC-lyn 6) en *levied at a standard rate of 14% on all goods* (KWIC-lyn 8).

Die term *statement of profit or loss and other comprehensive income* het 29 KWIC-lyne opgelewer. In twee van die KWIC-lyne is verklarende inligting vir die betrokke term gevind. Vergelyk in dié verband figuur 3.4 (vir die doel van die bespreking is die KWIC-lyne wat verklarende inligting bevat, geïsoleer):



Figuur 3.4. Geïsoleerde KWIC-lyne van die term *statement of profit or loss and other comprehensive income*.

Potensiële verklarende inligting wat bruikbaar is wat binne dié twee konkordansielyne gevind is, *is a statement that reflects the financial performance of an entity* en *the financial statement that indicates the profit and loss of an entity.*

Die verklarende inligting wat onttrek is kan nou gebruik word om terminologiese definisies saam te stel. Ter illustrasie word dié inligting gebruik om sodanige definisies saam te stel en word dit ook met bestaande definisies in ’n vakwoordeboek vergelyk.

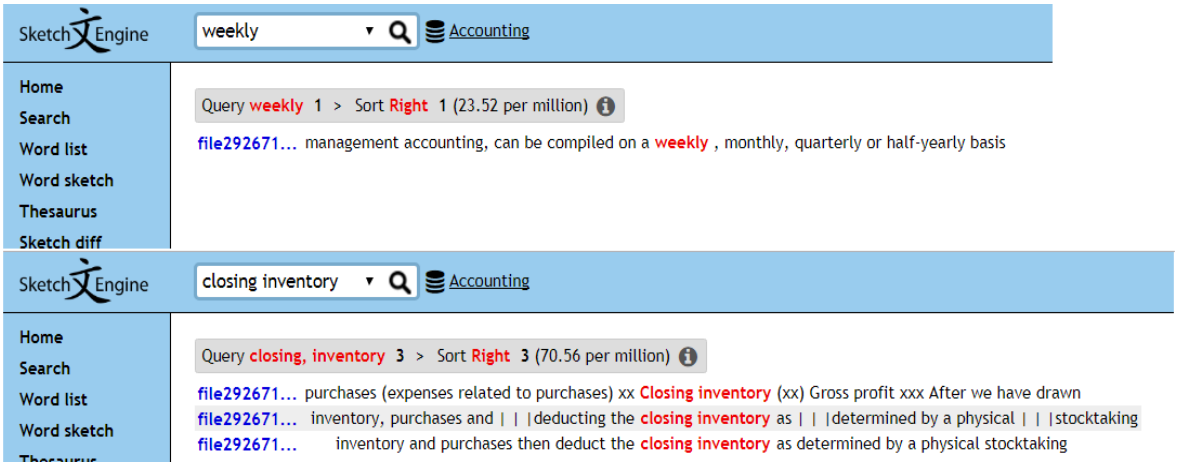
VAT (value-added tax) an indirect tax based on the consumption of goods and services typically by consumers (end-users). It is levied at a standard rate of 14%.

statement of profit or loss and other comprehensive income a statement that indicates the financial performance of an entity for a specific period, determining whether such an entity made a profit or a loss.

’n Bestaande definisie vir VAT (value-added tax) is: *an indirect tax on the consumption of goods and services in the economy* (www.sars.gov.za). ’n

Bestaande definisie vir die term *statement of profit or loss and other comprehensive income*, beskikbaar by www.accaglobal.com, is: *The performance of a company is reported in the statement of profit or loss and other comprehensive income.* Die terminologiese definisies wat saamgestel is, stem grootliks ooreen met bestaande definisies wat bevestig dat vakspesialiste bewustelik/onbewustelik sekere terme in vakspesifieke tekste verduidelik.

Dit is wel moontlik dat geen verklarende inligting vir ’n bepaalde term onttrek kan word nie – vergelyk figuur 3.5 hieronder vir KWIC-lyne vir die terme *weekly* en *closing inventory*:



Figuur 3.5. ’n Skermskoot van die terme wat deur die vakspesialiste bygevoeg is.

Die betrokke terme het nie aan die minimum frekwensiewaarde van vyf voldoen nie en is dus nie rekenaarmatig onttrek nie. Hieruit word afgelei dat dit ’n uitdaging is om verklarende inligting te onttrek vir terme wat nie aan die minimum frekwensiewaarde voldoen nie; weinig of geen verklarende inligting sal onttrek kan word nie. Nog ’n moontlike rede waarom geen verklarende inligting vir ’n bepaalde term onttrek kan word nie is dat sulke terme tot die kategorie van subtegniese woorde behoort. Leksikale items in dié kategorie is algemene woorde wat gespesialiseerde betekenisse aangeneem het in meer as een vakgebied en dus word die aanname gemaak dat hul betekenisse welbekend is (Pearson 1998:19); die term *weekly* is ‘n goeie voorbeeld van so ’n subtegniese item.

3.7 Samevatting

Die moderne benadering tot termidentifisering, wat met behulp van sagteware en elektroniese korpora rekenaarmatig uitgevoer word, lewer meer akkurate data op as die tradisionele benadering, omdat menslike subjektiwiteit tot ’n groot mate binne dié benadering uitgesluit word. Leksikale items wat beskou word as items met ’n lae frekwensie hoef nie slegs handmatig geëkserpeer word nie, want dit word wel semi-outomaties onttrek. Binne die onderhawige studie word *Sketch Engine* slegs op grond van die toegang tot die versameling algemene verwysingskorpora aanbeveel; ’n werklike evaluasie van *WordSmith Tools* en *Sketch Engine* sal ’n veel meer omvattender ondersoek vereis. Die moderne benadering maak nie net voorsiening dat enkel- en meerwoordterme en verklarende inligting onttrek word nie, maar ander nuttige data oor kollokasies, semantiek en sintaksis kan ook geïn word. Deur sekere terme bewustelik/onbewustelik te verduidelik wanneer vakspesifieke tekste saamgestel word, is dit moontlik om verklarende inligting semi-outomaties vanuit doelgemaakte korpora te onttrek. Hierdie inligting kan gebruik word om terminologiese definisies saam te stel.

HOOFSTUK 4

Gebruiksvoorbeelde en kollokatiewe inligting in ’n aanlyn termbank

4.1 Inleiding

Die voorsiening van ’n goeie gebruiksvoorbeeld aan die teikengebruiker is die laaste groot komponent wat in die leksikografiese proses voltooi moet word: “Finally … illustrating usage by examples. We no longer restrict ourselves to quoting from “the best writers” … but the provision of good illustrative examples is the last major component in the lexicographic process” (Rundell 2012:16). ’n Goeie gebruiksvoorbeeld stel die beoogde teikengebruiker in staat om terme, wat voorheen onbekend was, in konteks te plaas, bied linguistiese ondersteuning en bevorder die begrip van terme. In hierdie hoofstuk word daar oor die kenmerke van sodanige gebruiksvoorbeelde uitgebrei en ’n inventaris word gegee van Suid-Afrikaanse vakwoordeboeke wat gebruiksvoorbeelde bevat. Die behoeftes wat rondom gebruiksvoorbeelde binne die Suid-Afrikaanse konteks ontstaan, word bespreek. Die semi-outomatiese onttrekking van kollokatiewe inligting word ook behandel, aangesien die voorsiening van dié inligting aan Suid-Afrikaanse studente – die meerderheid wie se sterkste taal nie Engels is nie – onontbeerlik is (Nesselhauf 2003:22, Otto 1998:110). Daar word gekyk na die implikasies van die vertaling van kollokasies, asook na die moontlikhede oor hoe om kollokatiewe inligting binne ’n aanlyn termbank aan te bied. Laastens word die semi-outomatiese onttrekking van korpusgebaseerde gebruiksvoorbeelde met behulp van *Good Dictionary Example eXtractor* (afgekort as GDEX) (deel van die *Sketch Engine* suite) bespreek, asook die vertaling van gebruiksvoorbeelde.

4.2 Kenmerke van ’n goeie gebruiksvoorbeeld

Gebruiksvoorbeelde verrig verskillende funksies binne verskillende woordeboeke. Dié funksies word bepaal deur die tipe woordeboek en die teikengebruiker se profiel. Ongeag die funksie van die woordeboek en die teikengebruiker se profiel moet alle gebruiksvoorbeelde aan sekere vereistes voldoen. Atkins en Rundell (2008:459-461) brei uit oor die kenmerke van ’n goeie gebruiksvoorbeeld (sien paragraaf 1.5). Hulle gaan van die veronderstelling uit

dat gebruiksvoorbeelde vanuit ’n korpus onttrek word en meld dat ’n geskikte gebruiksvoorbeeld maklik herken kan word, want ’n (groot) korpus dui die konteks van ’n leksikale item, tipiese kollokasies, sintaktiese patrone en die meerwoord-uitdrukkings waarin die betrokke leksikale item gevind word, aan. Hierdie vier aspekte verteenwoordig ’n leksikale item se tiperende linguistiese eienskappe. Dit is ook belangrik dat ’n natuurlike gebruiksvoorbeeld ’n konstante register handhaaf. Vir ’n gebruiksvoorbeeld om informatief te wees, moet dit eerstens ’n duidelike funksie hê. Indien ’n gebruiksvoorbeeld van geen nut vir die teikengebruiker gaan wees nie, kan die spasie in ’n gedrukte woordeboek beter benut word. In die tweede plek moet die inligting in die gebruiksvoorbeeld nie té lank of té kort wees nie. Derdens moet die terminoloog ook verseker dat die inligting in die definisie nie teenstrydig met die gebruiksvoorbeeld is nie. Die teikengebruiker se profiel beïnvloed die keuse van gebruiksvoorbeelde; ’n gebruiksvoorbeeld kan natuurlik, kenmerkend én oorspronklik wees, maar indien die teikengebruiker dit nie begryp nie, is selfs die beste gebruiksvoorbeeld nutteloos. Atkins en Rundell (2008:461) beklemtoon ook die rol wat ’n korpus as bron van gebruiksvoorbeelde speel en meld dat ’n korpus wel gebruiksvoorbeelde kan oplewer wat aan die bogenoemde drie vereistes, d.i. natuurlikheid, kenmerkendheid en oorspronklikheid voldoen: “… if you use the corpus carefully and get the right balance between the three criteria … the examples you produce should bring real benefit for users.”

4.3 Inventaris van Suid-Afrikaanse vakwoordeboeke met gebruiksvoorbeelde

Binne die onderhawige studie is ’n inventaris saamgestel van al die kommersieel-beskikbare Suid-Afrikaanse vakwoordeboeke wat gebruiksvoorbeelde bevat (sien tabel 4.1). Die inventaris is saamgestel om vas te stel hoeveel vakwoordeboeke met gebruiksvoorbeelde beskikbaar is en watter tipe gebruiksvoorbeelde, hetsy frases of sinne, in die betrokke woordeboeke weergegee word. Die inventaris bevat woordeboeke wat in Suid-Afrika saamgestel is en woordeboeke wat gebaseer is op publikasies van die Verenigde State en Verenigde Koninkryk wat ook in Suid-Afrika beskikbaar is; talle Suid-Afrikaanse grondslagwoordeboeke is byvoorbeeld afkomstig van die

Verenigde Koninkryk (Hiles 2009:30). Die Suid-Afrikaanse teikengebruiker sal egter minder baat vind by woordeboeke wat in die Verenigde State en Verenigde Koninkryk gepubliseer is, omdat die inhoud dikwels kontekksspesifiek is. ’n Suid-Afrikaanse regstudent sal byvoorbeeld waarskynlik weining baat by ’n vakwoordeboek wat in die Verenigde State saamgestel is, aangesien die Amerikaanse regstelsel wesenlik van die Suid-Afrikaanse regstelsel verskil.

Dit is belangrik om te meld dat vakwoordeboeke wat ensiklopediese voorbeelde bevat, níé by die betrokke inventaris ingesluit is nie. Hierdie tipe voorbeelde dra semantiese inligting oor en vul die verklarende inligting aan, teenoor gebruiksvoorbeelde wat linguistiese inligting oordra. Ter illustrasie word die volgende ensiklopediese voorbeeld weergegee wat verkry is uit *A Dictionary of Grammatical Terms in Linguistics* (1993:4):

> **‘Bach Peters sentence’** One in which each of two noun phrases includes a pronoun anaphoric to the other: e.g. in *The girl who won it really deserves the prize she won,* the first pronoun (*it*) might have the same referent as *the prize she won,* the second (*she*) the same referent *as the girl who won it.*

In die voorbeeld hierbo is die sin *The girl who won it really deserves the prize she won* wel ’n voorbeeld van ’n *Bach Peters sentence*, maar dit maak duidelik deel uit van die ensiklopediese inligting wat verstrek word, en is nie primêr ’n voorbeeldsin wat die gebruik van die lemma illustreer nie.

Geïllustreerde vakwoordeboeke is ook nie by die inventaris ingesluit nie. Hoewel terme in geïllustreerde vakwoordeboeke deur middel van ’n figuur, formule, diagram of tabel uitgedruk word, bied dit nie aan die teikengebruiker die ondersteuning oor hóé om terme binne konteks te gebruik nie.

Met die samestelling van die onderstaande inventaris is daar vasgestel dat daar ook spesiale woordeboeke beskikbaar is wat vir spesifieke teikengebruikers saamgestel is en gebruiksvoorbeelde bevat, maar nie as vakwoordeboeke geklassifiseer word nie, byvoorbeeld: die tweetalige Pharos Mini-woordeboek bevat nuttige frases wat toeriste en besigheidsreisigers wat Suid-Afrika besoek, kan gebruik. Dit is ook belangrik om te noem dat hoewel daar gebruiksvoorbeelde in die vakwoordeboeke opgeneem is wat in die betrokke

inventaris gelys is, word elke term / lemma nie noodwendig van ’n gebruiksvoorbeeld vergesel nie. Dit is nie altyd duidelik op grond waarvan die terminoloog besluit het om ’n gebruiksvoorbeeld in te sluit nie. In sommige gevalle word ’n gebruiksvoorbeeld aangebied vir terme wat ’n lae frekwensie het: Frankenberg-Garcia (2012:286) dui aan: “... examples for decoding are more likely to help when the target word is a low frequency ... the context will probably contain words that are easier than the target word itself, facilitating comprehension.” Die betrokke inventaris bevestig dat dit eerstens nie algemene praktyk is om gebruiksvoorbeelde in gedrukte vakwoordeboeke in te sluit nie en tweedens, beklemtoon dit ook die tekort aan (Suid-Afrikaanse) vakwoordeboeke wat gebruiksvoorbeelde bevat, maar meer spesifiek dan, vakwoordeboeke waar élke term saam met ’n gebruiksvoorbeeld aangebied word. In aanlyn termbanke kan gebruiksvoorbeelde – en selfs meerdere gebruiksvoorbeelde – deel van die datakategorieë van elke term uitmaak, omdat ruimtebesparing in elektroniese termbanke nie ’n primêre oorweging is nie. Hoewel die terminoloog in teorie onbeperkte spasie binne ’n aanlyn termbank tot sy/haar beskikking het, moet daar gewaak word teen ’n oormaat inligting, die sogenaamde *information overload*; Rundell (2012:25) vat die onbeperkte spasie tot ’n terminoloog se beskikking binne ’n aanlyn termbank soos volg saam: “... this does not give editors carte blanche to include whatever data they want: notions of relevance and efficient information transfer still apply.”

Naam van woordeboek	**Teikengebruiker**	**Tipe gebruiksvoorbeelde**	**Datum van publikasie**
The Cambridge Mathematics Dictionary for Schools	graad 4 tot 9	sinne en illustrasies	2008
Die Cambridge Wiskunde-woordeboek vir skole (ook beskikbaar in: Engels, isiXhosa, Sesotho, Setswana, Siswati, isiNdebele, isiZulu en Noord-Sotho)	graad 4 tot 9	sinne	2008
Longman Multilingual Maths Dictionary (English, isiXhosa, Afrikaans)	graad 4 tot 6	sinne	2008
Longman Dictionary of Financial Terms	graad 10 tot 12, voor- en nagraadse studente	sinne	2007
Routledge Dictionary of Business	voor- en nagraadse studente	sinne	1999

Routledge Dictionary of Medicine	voor- en nagraadse studente, praktisyns	sinne	1999
Routledge Dictionary of Law	voor- en nagraadse studente	sinne	1999

Tabel 4.1. ’n Inventaris van Suid-Afrikaanse vakwoordeboeke met gebruiksvoorbeelde.

4.3.1 Gebruiksvoorbeelde binne die Suid-Afrikaanse konteks

In paragraaf 1.5 is daar vasgestel dat gebruikers binne die Suid-Afrikaanse konteks ’n behoefte het aan gebruiksvoorbeelde in die vorm van volledige sinne, in plaas van frases, aan gebruiksvoorbeelde wat van ’n aanvaarbare lengte is wat die begrip van ’n term sal bevorder en, waar moontlik, aan kollokatiewe inligting. Om in dié behoeftes te voorsien, word die volgende riglyne voorgestel:

- Tipe gebruiksvoorbeelde: gekose gebruiksvoorbeelde moet in die vorm van volledige sinne aangebied word; volledige sinne word geïdentifiseer deur ’n sin wat met ’n hoofletter begin en eindig met ’n leesteken soos ’n punt, vraagteken of uitroepteken.
- Lengte: ’n volledige sin van 10 tot 25 woorde dien as verstek vir die lengte van ’n gebruiksvoorbeeld. Indien die gebruiksvoorbeeld egter vertaal word, sal die lengte van die vertaalde gebruiksvoorbeeld verskil van dié van die oorspronklike gebruiksvoorbeeld wat uit die brontaal geselekteer is en kan so ’n voorbeeld buite die gestelde lengteparameters val.
- Kollokasies: volledige sinne wat kollokasies bevat, geniet voorkeur. Kollokasies hoef egter nie net in sinne (implisiet) aangebied word nie, maar kan ook apart (eksplisiet) aangebied word (sien paragraaf 4.8 vir ’n volledige bespreking).
- Sinne wat voornaamwoorde soos *hierdie*, *daardie* en *dit* bevat, asook anafore, moet liefs vermy word. Dikwels faal hierdie woorde om enige sinvolle betekenis oor te dra sonder enige verdere konteks (Kilgarriff et al. 2008:426).

Ten einde die toepassing van die riglyne te illustreer, is die volgende voorbeelde vanuit ’n lys van akademiese terme geneem (beskikbaar op oertb.tlterm.com). (Gebruiksvoorbeelde is in skuinsdruk.)

> Eng: **annual** (adj.) describing quantities or rates that are calculated over the period of a year. *The **annual income** of a top CEO in South Africa is more than one million rand.*

Eng: **fundamental** (adj.) absolutely necessary and important as a part of something, to the point that the second thing could not exist or succeed without it. *A supply of clean water is **fundamental to** the health of a nation.*

Eng: **view** (n) a personal opinion about something. *Convincing reasons must be provided in support of your **point of view** to increase the persuasiveness of the argument.*

Die bostaande gebruiksvoorbeelde is in die vorm van volledige sinne aangebied waar die lengte wissel tussen 13 en 19 woorde. Anafore en onbepaalde of besitlike voornaamwoorde word nie daarin gebruik nie. Kollokatiewe inligting word implisiet aangebied, dit wil sê binne die gebruiksvoorbeelde en nie as aparte datakategorieë nie. (Binne die betrokke studie is dit tipografies vir die leser in vetdruk aangedui.)

4.4 Kollokasies

Kollokasies, wat in die breë sin gedefinieer kan word as woorde wat dikwels naasmekaar in natuurlike taal geplaas word (Lin 1998:57, Kübler en Pecman 2012:188, McKeown en Dragomir 2000, Stubbs 1995:1), is binne die tradisionele benadering tot terminologie in gedrukte vakwoordeboeke slegs by wyse van uitsondering aan die teikengebruiker voorsien. Die mees voor die hand liggende rede hiervoor is natuurlik die eis om ruimte te bespaar. Nog ’n moontlike rede, soos wat Taljard (2016:553) daarop wys, is dat vakspesialiste – wat ’n belangrike rol in die samestelling van sodanige woordeboeke gespeel het – dié inligting as nutteloos beskou het. Taljard voeg by dat die hoofrede heel waarskynlik die beskouing is dat terme onafhanklik van enige konteks is (sien paragraaf 2.1) en dat kontekstuele inligting nie as belangrik geag word nie. Die fokus het egter die afgelope jare verskuif na aanlyn termbanke, waar die eis om ruimtebesparing verval en die moderne benadering tot terminologie maak voorsiening vir die bestudering van terme binne konteks (sien paragraaf 2.3).

Volgens Evert (2007:2) word daar onderskeid getref tussen empiriese en fraseologiese (Eng: *phraseological*) kollokasies. Hy definieer empiriese kollokasies as “... recurrent and predictable word combinations, which are a directly observable property of natural language …” Voorbeelde van sodanige

kollokasies is *koffie* en *tee*, *raad* en *daad* en *hart* en *siel*. Fraseologiese kollokasies, ook bekend as meerwoord-uitdrukkings, word gedefinieer as: "… semi-compositional and lexically determined word combinations …" (Evert 2007:2). Voorbeelde van hierdie tipe kollokasies is *blakend gesond, geduldig wag* en *bose kringloop*. Evert voeg by dat fraseologiese kollokasies in subkategorieë verdeel word: "… subcategories ranging from completely opaque idioms to semantically compositionally word combinations, which are merely subject to arbitrary lexical restrictions …" Afhangende van die teikengebruiker, kan albei tipes kollokasies van nut wees. Taljard (2016:554) meld dat 'n databasis die bron van data vir 'n verskeidenheid van terminologiese toepassings met verskillende funksies kan wees, byvoorbeeld kognisie, teksproduksie, teksresepsie en vertaling. Sy meld dat empiriese kollokasies relevant is op 'n konseptuele vlak, aangesien die samestellende dele van 'n kollokasie konseptueel met mekaar verband mag hou, en binne die funksie van kognisie is hierdie inligting nuttig. Aan die ander kant, voeg Taljard by, is fraseologiese kollokasies op 'n pragmatiese vlak relevant, want dit hou verband met die gebruiksvlak; dit is belangrik binne die funksies van teksproduksie en vertaling.

Kübler en Pecman (2012:201) meld dat die terminoloog twee tipes fraseologiese data in 'n databasis kan insluit: algemene kollokasies (ook bekend as term-onafhanklike kollokasies) en vakspesifieke kollokasies (ook bekend as term-afhanklike kollokasies). Dié kollokasies kan handmatig of semi-outomaties onttrek word. Indien eersgenoemde benadering gevolg word, sal die terminoloog nie rekenskap kan gee van al die kollokasies wat binne 'n bepaalde korpus voorkom nie en kan die terminoloog ook nie akkurate data oor die frekwensie en verspreiding van verskillende kollokasies weergee nie (Stubbs 1995:2).

Daar is tans 'n verskeidenheid van sagteware, soos *WordSmith Tools*, beskikbaar wat gebruik kan word om inligting oor kollokasies semi-outomaties te onttrek (Kilgarriff en Kosem 2012:41). Dié inligting word onttrek volgens 'n minimum frekwensiewaarde wat deur die terminoloog gestel is en die data word volgens frekwensie gerangskik. Kilgarriff en Kosem (2012:41) wys daarop dat daar nie noodwendig 'n probleem is met die data wat met behulp van byvoorbeeld

WordSmith Tools onttrek word nie, maar dat die aanbieding van die data lomp is en die bestudering daarvan baie tyd in beslag neem. (Vergelyk in hierdie verband figuur 4.1 wat geneem is uit Kilgarriff en Kosem (2012:14) wat illustreer dat *WordSmith Tools* kollokatiewe inligting reëlgewys aanbied en dat die betrokke inligting volgens frekwensie gerangskik word.) Hulle voeg by dat hierdie metode grammaties blind is en dat dit net nabyheid (Eng: *proximity*) in ag neem.

TABLE 3.1. Top fifteen collocates of the verb *save* (ordered by MI score)

Lemma	freq[7]	MI
BuyerZone.com	7	13.19
ac);	5	13.19
count-prescription	5	13.19
Christ-A-Thon	7	13.19
Teldar	6	12.61
Re:What	26	12.55
Redjeson	5	12.51
INFOPACKETS30	3	12.46
other-I	4	12.39
SetInfo	4	12.39
Ctrl-W	9	12.36
God	18	12.23
Walnuttree	3	12.19
Hausteen	5	12.19
MWhs	3	12.19

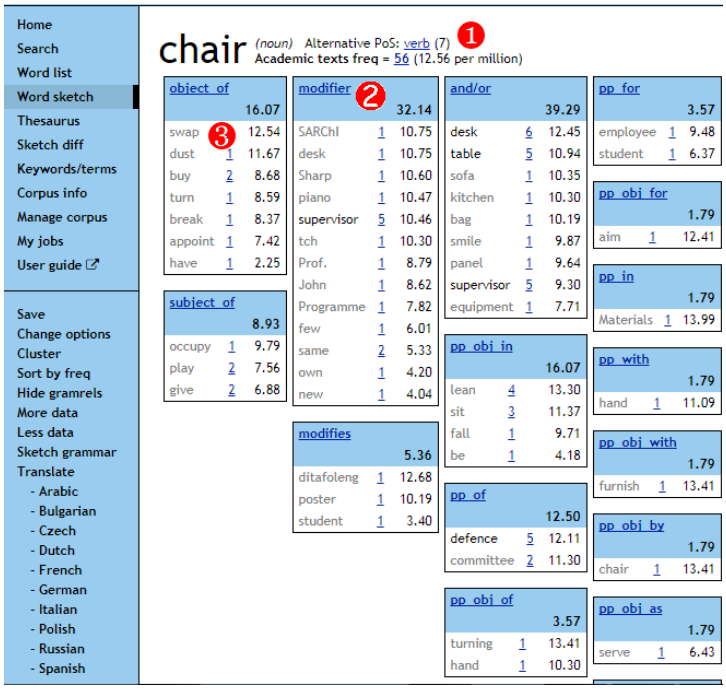
[7] WordSmith Tools lists collocates in the 'Patterns' view by frequency only.

Figuur 4.1. 'n Voorstelling van *WordSmith Tools* wat kollokatiewe inligting reëlgewys volgens frekwensie aanbied.

Om dié redes stel die betrokke studie ondersoek in na die semi-outomatiese onttrekking van kollokatiewe inligting vanuit doelgemaakte korpora deur middel van *Sketch Engine* se *Word Sketch*-funksie. Kilgarriff en Rundell (2002:807) beskryf die voordele van *Word Sketches* soos volg: "... Word Sketches not only streamlined the process of searching for significant word combinations, but often provided a more revealing, and more efficient way of uncovering the key features of a word's behaviour than the (now traditional) method of scanning concordances." Die *Word Sketch*-funksie word in die volgende paragrawe bespreek.

4.5 *Word Sketches*

Word Sketches analiseer korpora outomaties en lewer onmiddellike resultate. Die eindproduk is ’n opsomming in die vorm van een bladsy (in teenstelling met ander sagteware wat data reëlgewys volgens frekwensie rangskik) wat die grammatiese inligting en kollokasies van ’n bepaalde term bevat (Killgarriff et al. 2004:105, Killgarriff et al. 2010:372, www.sketchengine.co.uk). Soos wat dit deurentyd binne die betrokke studie beklemtoon word, speel die aard, omvang en samestelling van die korpus ’n belangrike rol in die data wat gelewer word - Kilgarriff et al. (2010:375) en Srdanović et al. (2011:72) meld dat die kwaliteit van die *Word Sketches* beïnvloed word deur die kwaliteit én grootte van die korpus, asook deur ander aspekte soos o.a. woordsoortannotering, grammatika en statistiek, aangesien hierdie komponente implisiet gebruik word om *Word Sketches* saam te stel. Die tipiese inligting wat in ’n *Word Sketch* verskyn, word in figuur 4.2 geïllustreer:



Figuur 4.2. Tipiese inligting wat in ’n *Word Sketch* weergegee word.

’n *Word Sketch* van die term *chair* as ’n selfstandige naamwoord, is gegenereer. Die korpus bestaan uit ’n versameling akademiese tekste. ’n *Word Sketch* gee die volgende inligting weer:

1. Die woordsoort van die soekwoord word tussen hakies aangedui, asook die alternatiewe woordsoort van ’n betrokke term. Die aantal kere wat die term as ’n spesifieke woordsoort voorkom, word ook aangedui. Binne die betrokke *Word Sketch* kom die term *chair* 56 keer as ’n selfstandige naamwoord en 7 keer as ’n werkwoord voor.
2. ’n Analise van die belangrikste sintaktiese eienskappe van die betrokke soekwoord word aangedui. In die betrokke geval is dit duidelik dat die bepalers *SARChI* en *desk* die hoogste kollokasiewaarde ten opsigte van *chair* het.
3. Die frekwensie van die kollokatiewe inligting word ook aangedui.

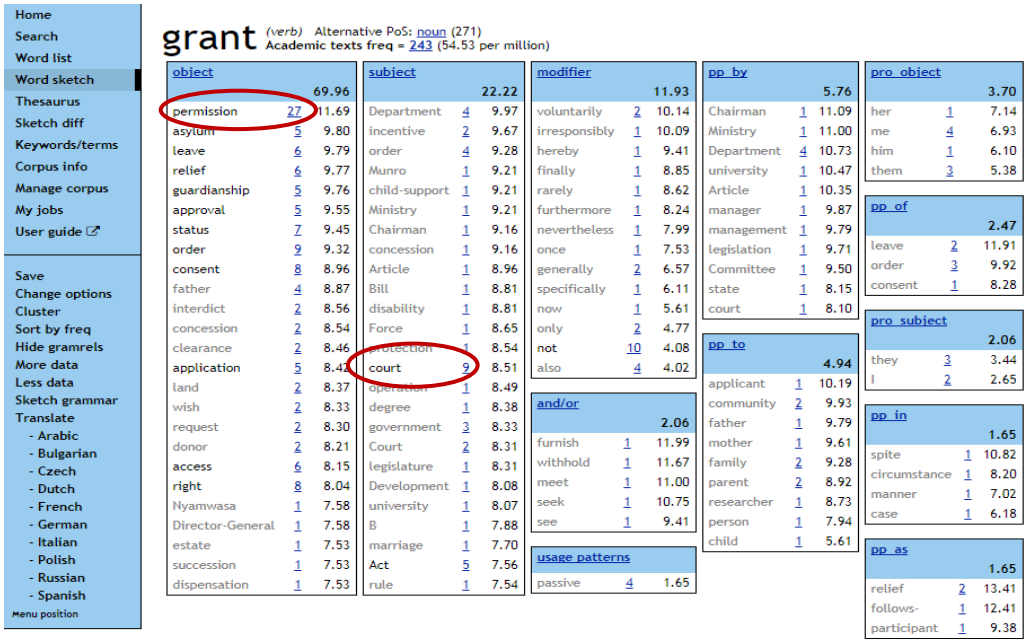
4.6 *Word Sketch* binne *Sketch Engine*

Die funksie *Word Sketch* binne die sagteware *Sketch Engine* is uniek, want: “Sketch Engine … takes as input a corpus of any language (with appropriate linguistic markup), and … generates, amongst other things, word sketches for the words of that language” (Kilgarriff et al. 2004:105). Dit het ook ander nuttige funksies, soos *Word Sketch Difference* (ook bekend as *sketchdiffs*), *Bilingual Word Sketch* en ’n tesourus. Elkeen van hierdie funksies word vervolgens kortliks bespreek.

Die terminoloog kan *Word Sketch Difference* gebruik om twee terme met mekaar te vergelyk om die ooreenkomste en verskille in kollokasies vas te stel (Kilgarriff en Kosem 2012:46). Tweedens kan *Bilingual Word Sketch* aangewend word om die woordsketse van ’n betrokke term en sy vertaling te bestudeer om te bepaal of hulle dieselfde betekenis het en of hulle kollokasies in gemeen het (www.sketchengine.co.uk). Laastens voorsien die tesourus die terminoloog met ’n lys woorde van die “naaste bure”, d.i. die meeste gemeenskaplike woorde wat saamgegroepeer word (Kilgarriff en Kosem 2012:45).

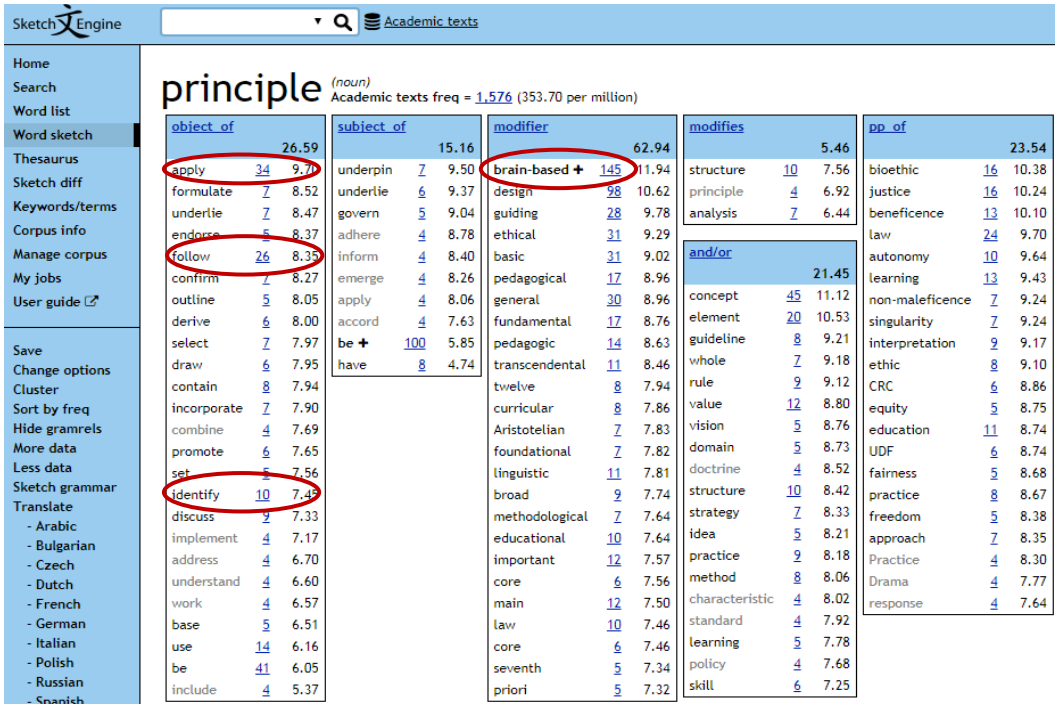
Om die *Word Sketch*-funksie te illustreer, is drie terme lukraak vanuit die akademiese termlys van die projek, wat 1 153 terme bevat, gekies: *grant* (werkwoord), *principle* (selfstandige naamwoord) en *random* (byvoeglike naamwoord). Die kollokatiewe data is uit ’n korpus van akademiese skryfwerk onttrek wat uit 3 582 315 woorde (*tokens*) bestaan.

Eerstens, kom die term *grant* (sien figuur 4.2) 243 keer as ’n werkwoord in die korpus voor. Die item wat die meeste kere saam met *grant* in die korpus voorkom, is *permission,* wat as objek van die oorganklike werkwoord *grant* gebruik word. Die item *court* word in 9 gevalle as onderwerp van die werkwoord *grant* gebruik.



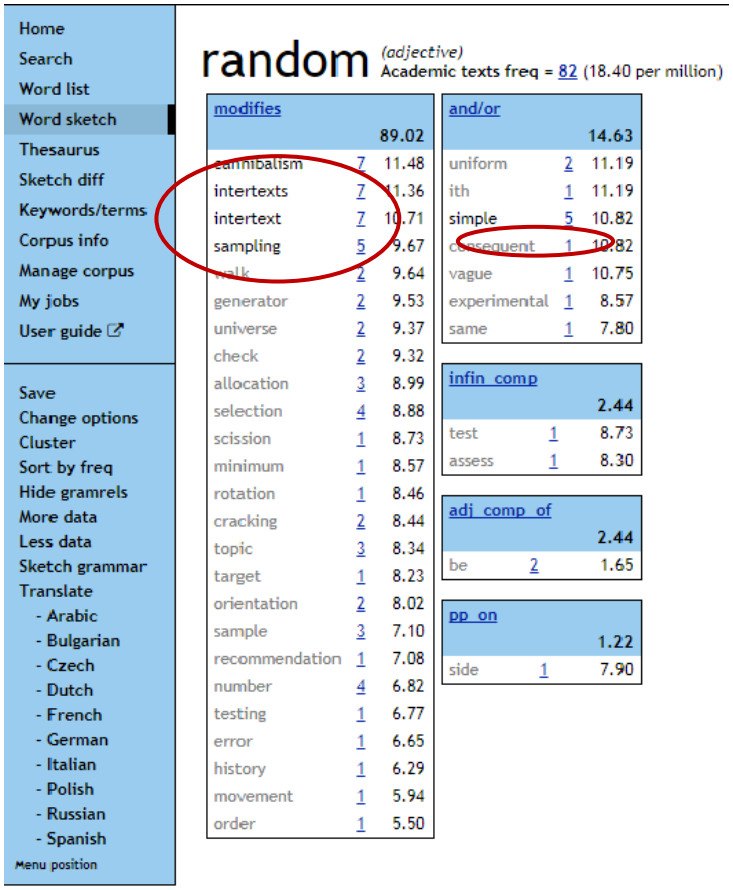
Figuur 4.3. Kollokatiewe inligting vir die term *grant*.

Tweedens, binne die akademiese korpus verskyn die term *principle* (sien figuur 4.3) 1 576 keer as ’n selfstandige naamwoord. Dit word oorwegend as voorwerp van oorganklike werkwoorde gebruik. Die werkwoorde waarmee saam dit dikwels voorkom, is *apply*, *follow* en *identify*. Die bepaler wat die meeste vir *principle* gebruik word, is *brain-based*.



Figuur 4.4. Kollokatiewe inligting vir die term *principle.*

Laastens, die term *random* (sien figuur 4.4) word binne die betrokke korpus 82 keer as ’n byvoeglike naamwoord gebruik. Slegs vier terme word as beduidende kollokasies deur die sagteware uitgewys, d.i. *cannibalism, intertexts, intertext* en *sampling* is naamwoorde wat deur die term *random* bepaal word en *simple* is vyf keer as en/of naas *random* geplaas. Lin (1998:57) meld egter dat hoewel kollokasies herhalend is, dit nie noodwendig ’n groot aantal keer in ’n korpus verskyn nie: “In order to achieve broad-coverage, a collocation needs to be extracted even if it only occurs a few times in the corpus.”



Figuur 4.5. Kollokatiewe inligting vir die term *random.*

In paragraaf 4.9 hieronder word ’n bespreking gegee van die wyse waarop hierdie kollokatiewe inligting deur die terminoloog benut kan word en tot die waarde van enige terminologiese produk kan bydra.

4.7 Vertaling van kollokasies

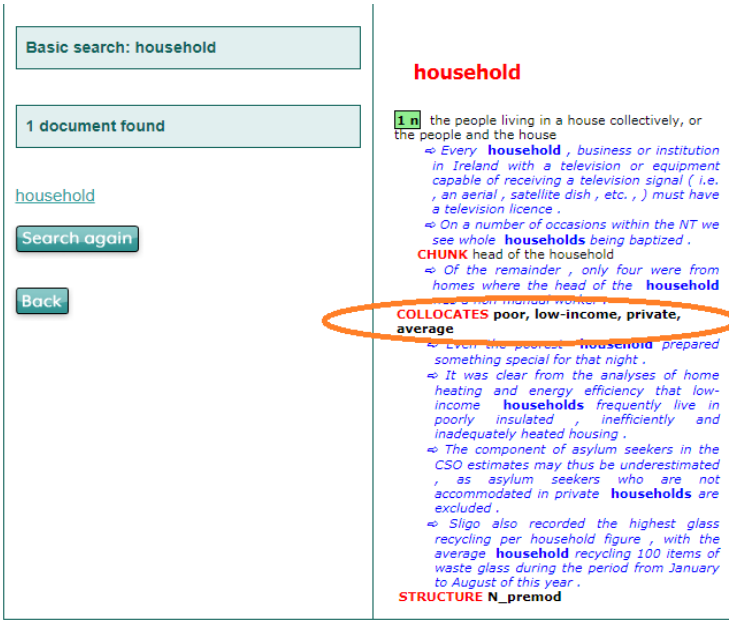
Binne die Suid-Afrikaanse konteks is Engels gewoonlik die brontaal waaruit definisies, gebruiksvoorbeelde, kollokasies, ens. na ’n ander amptelike taal of tale vertaal moet word. Binne die projek wat die fokus van die studie is, is daar bevind dat wanneer kollokasies vanaf Engels na die tien amptelike Afrikatale vertaal word, die betekenis van die kollokasies somtyds platval. L’Homme et al. (2012:216) meld dat kollokasies ’n onvoorspelbare kombinasie van leksikale eenhede is en dié kombinasie kan nie op grond van die gewone semantiese of sintaktiese eienskappe van die betrokke eenhede geproduseer word nie. ’n Moontlike rede waarom die vertaling van kollokasies problematies kan wees, is dat kollokasies taalspesifiek (McKeown en Dragomir 1997:16) en uniek is (Taljard

2016:554) en dus kan daar nie net aangeneem word dat die betekenis van ’n kollokasie in ’n betrokke taal dieselfde sal wees as in ’n ander taal nie. In Afrikaans byvoorbeeld, *stel* iemand ’n kontrak *op*, maar in Engels word die werkwoord *write up* gebruik, en nié *set up* nié. Nog ’n moontlike rede is dat die vertalers heel waarskynlik nie vakspesialiste is nie en nog nie die fraseologie van ’n bepaalde vakgebied bemeester het nie (Kübler en Pecman 2012:188). Kollokasies moet egter nie net bloot vertaal word nie - kollokatiewe ekwivalente wat vanuit die korpus van die teikentaal onttrek is, moet aan die teikengebruiker voorsien word (L’Homme et al. 2012:231, Kübler en Pecman 2012:202). Binne die Suid-Afrikaanse konteks is dit egter nie moontlik om ekwivalente binne vakspesifieke tekste in die Afrikatale te identifiseer nie, omdat daar weinig of geen van hierdie tipe tekste beskikbaar is nie. Dit sou dus baie moeilik wees om kollokatiewe inligting, anders as die wat uit die teikentaal na die Afrikataal as brontaal vertaal is, aan die teikengebruiker te voorsien.

4.8 Aanbieding en toegang tot kollokatiewe inligting binne ’n aanlyn termbank

Binne ’n aanlyn termbank kan kollokatiewe inligting op twee maniere aan die teikengebruiker beskikbaar gestel word: implisiet en eksplisiet. Kollokatiewe data wat vanuit (doelgemaakte) korpora onttrek is, is aanvanklik slegs van waarde vir die terminoloog. Die terminoloog sal dikwels hierdie data gebruik om gebruiksvoorbeelde (handmatig) te selekteer wat sover as moontlik kollokatiewe inligting bevat (Taljard 2015:399). Op dié wyse word kollokatiewe inligting op implisiete wyse aan die teikengebruiker oorgedra – dit verskyn dus nie as ’n aparte, diskrete datakategorie nie. Binne dié metode moet kollokatiewe inligting met voldoende toeligting aangebied word om die teikengebruiker in staat te stel om sodanige inligting te identifiseer. Daar moet dus besluit word hóé die betrokke kollokatiewe inligting binne die gebruiksvoorbeeld tipografies aangedui gaan word, hetsy in skuins- of vetdruk (sien paragraaf 4.3.1). Die terminoloog kan verder ’n gebruiksnota byvoeg met addisionele inligting oor byvoorbeeld die semantiese prosodie van die betrokke kollokatiewe inligting. Die terminoloog kan dus binne die gebruiksnota aandui of die kollokatiewe inligting ’n positiewe of negatiewe konnotasie het. Daarteenoor is die eksplisiete aanbieding van

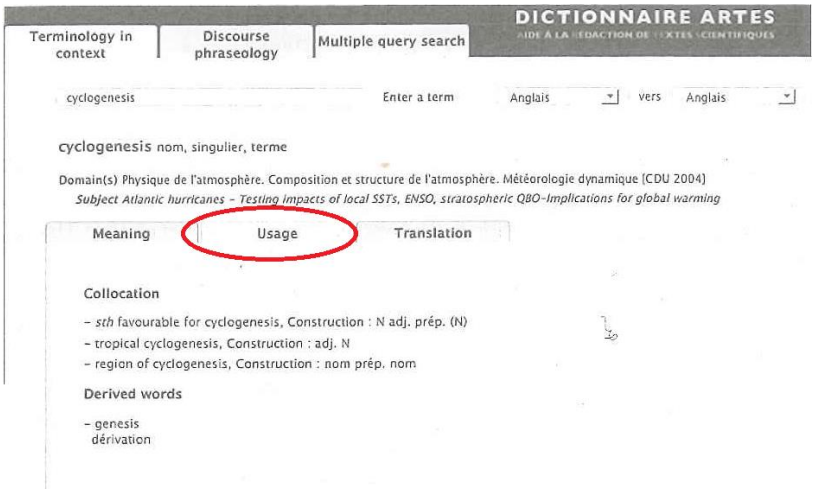
kollokatiewe inligting as aparte datakategorie wat vir die teikengebruiker duidelik onderskeibaar. Dit kan beteken dat kollokatiewe inligting outomaties as deel van die aanvanklike soekresultate aan die gebruiker beskikbaar gestel word. Taljard (2015:400) meld dat die terminoloog versigtig moet oorweeg oor hoe die kollokatiewe inligting binne dié metode aan die teikengebruiker ten toon gestel gaan word: word dit by verstek vertoon sodra die gebruiker 'n soektog uitvoer, of word die teikengebruiker die opsie gegee om deur middel van 'n gevorderde soektog toegang tot kollokatiewe inligting te verkry? Vergelyk in dié verband figuur 4.3 afkomstig van die *DANTE*-databasis (http://www.webdante.com/the_database.html) waar kollokatiewe inligting by verstek as resultaat van die aanvanklike soektog aan die teikengebruiker vertoon word. (In die betrokke figuur is die woord *household*, wat lukraak uit die sosiologielys van die projek gekies is, as soekwoord gebruik.)



Figuur 4.6. 'n Skermskoot van die *DANTE*-databasis se aanbieding van kollokatiewe inligting.

Die bostaande figuur dui aan dat die kollokatiewe inligting op die *DANTE*-databasis by verstek aan die teikengebruiker ten toon gestel word. Teikengebruikers moet eers deur die definisie en gebruiksvoorbeelde sif, alvorens hulle die kollokatiewe inligting kan raadpleeg. Taljard wys egter daarop dat die terminoloog daarteen moet waak om nie die teikengebruiker met té veel inligting te bombardeer nie: "*Collocational information should therefore be made available*

as an optional, additional search function which can be accessed by means of clicking on a dedicated button or tab." Die teikengebruiker sal dus meer baat vind by die eksplisiete aanbieding van sodanige inligting wat geraadpleeg kan word deur 'n toegekende skakel wat die teikengebruiker in staat stel om na spesifieke kollokasies te soek (L'Homme 2012:226). Vergelyk in dié verband figuur 4.4 wat geneem is uit Kübler en Pecman (2012:197):



Figuur 4.7. 'n Voorstelling van die *ARTES*-databasis waar kollokatiewe inligting eksplisiet aangebied word.

Die bostaande figuur illustreer dat kollokatiewe inligting binne die betrokke databasis d.i. *ARTES*-databasis eksplisiet ten toon gestel word en dat die teikengebruiker na spesifieke kollokatiewe inligting kan soek deur middel van 'n toegekende skakel – in hierdie geval is die toegekende skakel *Usage* benoem.

4.9 Korpusgebaseerde gebruiksvoorbeelde

Reeds in paragraaf 1.5 is daar aangedui dat kenners ten gunste is van korpusgebaseerde gebruiksvoorbeelde. Sodanige gebruiksvoorbeelde kan handmatig of semi-outomaties onttrek word. Kilgarriff et al. (2008:426) dui aan dat dit 'n uitdaging kan wees om iewers binne groot (algemene) korpora gebruiksvoorbeelde in die vorm van volledige sinne te vind wat aan die gestelde kriteria van 'n bepaalde projek voldoen. Die gebruiksvoorbeelde wat onttrek word, moet dikwels aangepas word, dit sluit in: irrelevante inligting wat verwyder moet word en/of 'n ingewikkelde woord wat met 'n eenvoudiger een vervang moet word. Binne doelgemaakte korpora sal dit dus 'n groter uitdaging wees om

gebruiksvoorbeelde te onttrek, aangesien dié tipe korpora dikwels kleiner is as algemene korpora. Die proses sal egter vergemaklik word indien gebruiksvoorbeelde semi-outomaties onttrek word, daarom stel die onderhawige studie ondersoek in na die semi-outomatiese onttrekking van gebruiksvoorbeelde vanuit doelgemaakte korpora met behulp van die funksie GDEX binne die *Sketch Engine*-sagteware suite.

4.10 GDEX

Good Dictionary Example eXtractor (afgekort as GDEX) is ’n sisteem wat sinne rekenaarmatig evalueer op grond van hul gepastheid om as gebruiksvoorbeelde te dien. Dié proses word outomaties uitgevoer: goeie gebruiksvoorbeelde word van swak gebruiksvoorbeelde geskei op grond van voorafgestelde kriteria. Die eindproduk is ’n gerangskikte lys van sinne waarin die terminoloog heel waarskynlik ’n goeie, gepaste gebruiksvoorbeeld sal vind. Kilgarriff en Rychlý (2010:412) voeg by dat daar nie net aangeneem kan word dat die “beste” gebruiksvoorbeeld volgens GDEX goed genoeg is om dadelik aan die teikengebruiker verskaf te word nie - die GDEX-funksie verbeter wel die kanse dat die terminoloog ’n bruikbare sin binne die eerste paar konkordansielyne vind. Dit is belangrik om te noem dat die terminoloog nie die konkordansies self moet sorteer aan byvoorbeeld die linker- of regterkant nie. Die beste gebruiksvoorbeelde volgens GDEX sal dan uiteraard nie binne die eerste paar konkordansies voorkom nie, Kilgarriff et al. (2008:430) beweer in dié verband: “*…* non-grammatical and ‘junk’ corpus examples are tucked away towards the end of the concordances so [they] are not shown to the lexicographer unless they scroll through hundreds of examples.” Sekere aspekte van GDEX moet wel nog aangepas word, voordat GDEX sy optimale potensiaal bereik (Kilgarriff et al. 2008:431, Rundell 2012:28). In die eerste plek moet metodes geïmplementeer word om die sinne wat onttrek word en as “gemors” beskou word, uit te wys. Tweedens moet daar vasgestel word of die outomatiese sintaksontleder moeilike sinne as swak gebruiksvoorbeelde klassifiseer (Kilgarriff et al. 2008:429). Afgesien hiervan stel die GDEX-funksie nog steeds die terminoloog in staat om gebruiksvoorbeelde vanuit doelgemaakte korpora aan die teikengebruiker te

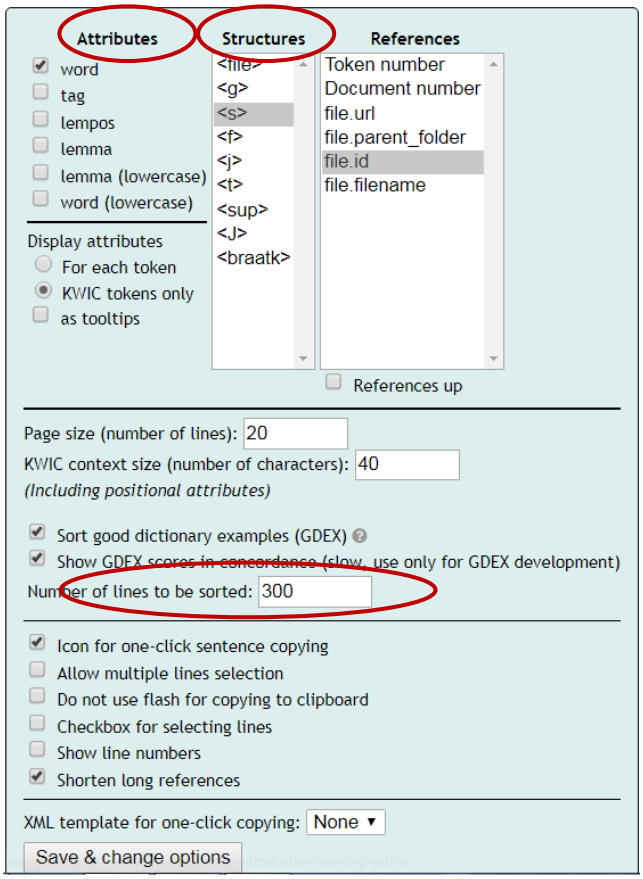
voorsien wat ’n ware weerspieëling van egte, “lewende” taal is (sien paragraaf 1.5).

4.11 GDEX binne *Sketch Engine*

Binne die sagteware *Sketch Engine* word GDEX gebruik om gebruiksvoorbeelde volgens twee tegnieke te onttrek: die sortering van konkordansielyne en die algoritme-gebaseerde metode (www.sketchengine.co.uk). Elkeen van hierdie tegnieke word vervolgens bespreek.

Om geskikte gebruiksvoorbeelde vir ’n bepaalde term volgens die eersgenoemde tegniek te soek, word sinne in konkordansielyne gesorteer. Die “beste” gebruiksvoorbeelde volgens GDEX sal dus in die eerste paar konkordansielyne voorkom. Dié funksie word slegs geaktiveer wanneer die terminoloog die kriteria vir ’n bepaalde soektog bepaal het. Die kriteria word beïnvloed deur die volgende aspekte (www.sketchengine.co.uk) (sien figuur 4.8):

- Kenmerke: addisionele inligting wat verband hou met elke token in ’n korpus. Hierdie addisionele inligting is gewoonlik versteek. Die terminoloog moet die kenmerke wat hy/sy in die konkordansie wil hê, merk. Daar word vereis dat ten minste één kenmerk gemerk word. Die kenmerke waarvan die terminoloog kan kies, sluit in: *woord, merker, lempos* (’n kombinasie van die lemma en woordsoort), *lemma (kleinletter)* en *woord (kleinletter)*.
- Strukture: verwys na die segmente waarin ’n korpus verdeel kan word, byvoorbeeld sinne en paragrawe. Dié aspek bepaal die strukture wat die terminoloog in die konkordansie wil sien.
- Aantal konkordansielyne vir sortering: die terminoloog bepaal die aantal lukrake sinne wat volgens GDEX sorteer moet word; 300 sinne word as verstek gebruik.



Figuur 4.8. ’n Skermskoot van die kriteria wat gestel word om gebruiksvoorbeelde in konkordansielyne te sorteer.

Ter illustrasie is drie terme lukraak vanuit die akademiese lys van die projek gekies: *ambiguous* (byvoeglike naamwoord), *convince* (werkwoord) en *rhythm* (selfstandige naamwoord). Die konkordansies is volgens die kenmerk *woord* sorteer en die gekose struktuur binne die konkordansielyne is sinne. Driehonderd lukrake sinne word deur GDEX sorteer. Hoewel GDEX die “beste” gebruiksvoorbeelde eerste sorteer, rus die onus nog steeds op die terminoloog om die konkordansielyne te bestudeer en te besluit watter sinne die mees gepaste gebruiksvoorbeelde is. Vir die doeleindes van dié klein eksperiment, moet ’n gebruiksvoorbeeld natuurlik, kenmerkend én informatief wees om as ’n goeie, bruikbare gebruiksvoorbeeld te kwalifiseer.

Die eerste term *ambiguous* het 27 konkordansies opgelewer. Binne die eerste tien konkordansies is ses bruikbare gebruiksvoorbeelde gevind. Die res van die konkordansies het drie bruikbare gebruiksvoorbeelde bevat. Die tweede term *convince* het 72 konkordansies opgelewer; die eerste tien konkordansies bevat

agt bruikbare gebruiksvoorbeelde. In die res van die konkordansies is nog 13 sinne gevind wat as gebruiksvoorbeelde kan dien.

Die derde term *rhythm* het 81 konkordansies opgelewer. Binne die eerste tien konkordansies is een bruikbare gebruiksvoorbeeld gevind; ses sinne verwys na die meerwoord-term *circadian rhythm* en kan dus nie binne die betrokke konteks gebruik word nie. Die res van die konkordansies het slegs drie gebruiksvoorbeelde opgelewer.

Hoewel dit nie die doel van die klein eksperiment was om vas te stel hoeveel gebruiksvoorbeelde in die konkordansies voorkom nie, was dit interessant om waar te neem en te bevestig dat die GDEX-funksie wel binne die eerste paar konkordansies die mees bruikbare gebruiksvoorbeelde sorteer. Daar is ook vasgestel dat van hierdie gebruiksvoorbeelde aangepas moet word, voordat dit aan die teikengebruiker voorsien word. Die term *rhythm* het wel nie belowende resultate opgelewer nie. Dit word egter nie toegeskryf aan die tekortkominge van die GDEX-funksie nie, maar eerder aan die inhoud en grootte van die korpus.

Die gekose konkordansielyne wat gebruiksvoorbeelde bevat kan ook gekopieer word om in ander toepassingsagteware gebruik te word, mits die korpus die gekose formaat ondersteun (Rundell 2012:22). Kilgarriff en Kosem (2012:50) meld: “Copy-and-paste is possible in some cases, but often the information needs to be in a specific format (normally XML) ...” Binne die projek is dit egter nie moontlik om die data te kopieer en in die databasis te plak nie, aangesien die korpora nie die formaat ondersteun nie.

Deur *Word Sketch* en GDEX te kombineer, maak die algoritme-gebaseerde metode gebruik van *TickBox Lexicography* (afgekort as TBL) om gebruiksvoorbeelde te onttrek. Die proses word uitgevoer deur *Word Sketch* wat ’n kandidaatlys met kollokatiewe inligting van ’n betrokke term oplewer; hieruit merk die terminoloog die kollokasies wat hy/sy in die termbank wil insluit. Vir elke kollokasie wat gemerk is, lewer GDEX ses gebruiksvoorbeelde op. Binne die onderhawige studie kon geen gebruiksvoorbeelde deur middel van TBL onttrek word nie, omdat dié funksie nie by verstek by ’n gebruiker se subskripsie aan *Sketch Engine* ingesluit is nie (www.sketchengine.co.uk). Kundiges is wel ten

gunste van die gebruik van TBL en die vooruitsig is dat die terminoloog ’n ander rol sal aanneem: TBL sal goeie, gepaste gebruiksvoorbeelde selekteer en die terminoloog sal die gekose gebruiksvoorbeelde evalueer en redigeer (Rundell 2012:23). Kilgarriff et al. (2010:418) voeg by: “... TBL has great potential for both streamlining corpus lexicography and making it more accountable to the corpus.”

In teenstelling met die bestudering van konkordansielyne, dra TBL inligting direk oor na databasisse (Rundell 2012:23). Volgens Kilgarriff en Kosem (2012:50) word ’n XML-modelvorm benodig om die data oor te dra in die formaat wat versoenbaar is met die betrokke databasis, maar voeg by: “The lexicographer does not need to think about XML: from their perspective, it is a simple matter of copy-and-paste.” Die TBL-toepassing gebruik dus dieselfde datakategoriebenaminge as wat algemeen in databasisse voorkom, byvoorbeeld *lemma, definisie, gebruiksvoorbeeld,* en hoef die terminoloog slegs die data in die betrokke databasis te plak.

Benewens dié tegniek, kan die terminoloog ook gebruik maak van sagteware soos *TLex Dictionary Compilation Software* (Joffe en de Schryver 2004, Rundell 2012:25) waar die korpus-funksie en die databasis gekombineer is. Dit stel die terminoloog in staat om dieselfde koppelvlak te gebruik, d.i. binne die korpus te soek én inskrywings in die databasis saam te stel. Verder kan die betrokke sagteware ook die lengte van gebruiksvoorbeelde monitor (Rundell 2012:25), wat verseker dat die terminoloog by die voorafgestelde kriteria hou. *TLex Dictionary Compilation Software* is reeds geskik om binne die Suid-Afrikaanse konteks gebruik te word. Die nadele egter verbonde aan hierdie tipe sagteware, soos wat Kosem (2016:91) daarop wys, is dat “an integrated corpus query tool can never be as sophisticated as standalone software solutions, and cannot be used on larger (reference) corpora without compromising the functionality of the dictionary-writing part of the software.”

4.12 Vertaling van gebruiksvoorbeelde

Die vertaling van gebruiksvoorbeelde vanuit ’n brontaal na ’n teikentaal is nie sonder probleme nie, veral in gevalle waar die teikentaal ’n Afrikataal is. Voorbeelde uit die praktyk het aan die lig gebring dat die term wat deur die

gebruiksvoorbeeld geïllustreer moet word, dikwels in die vertaling verlore gaan. Die vertaalde terme word dus eerstens nie binne konteks geplaas nie, tweedens word die begrip van die betrokke vertaalde terme nie bevorder nie, en laastens bied sodanige gebruiksvoorbeelde nie die nodige linguistiese ondersteuning aan die teikengebruiker nie. Ter illustrasie is die volgende drie gebruiksvoorbeelde lukraak vanuit die Sesotho-vertaling van die akademiese lys onttrek. 'n Sesotho-kwaliteitskontroleerder is gevra om die vertalings na te gaan en die probleme, indien enige, uit te wys. Die brontaal, Engels, word telkens saam met die vertaling weergegee.

Eng: **exhibit** (v) to show, make visible or display. *The artist's paintings will be exhibited at the re-opening of the art gallery.*

Ses: **totobatsa** ho bontsha, ho etsa hore ho bonahale kapa ho pepesa. *Ditshwantsho tsa setsebi sa bonono di tla pepeswa ha ho kgakolwa sebaka sa dipontsho tsa bonono*.

Eng: **significant** (adj.) important in effect or meaning. *The increase in the petrol price will have a significant effect on food prices.*

Ses: **e kgolo** e kgolo ka sekgahla kapa moelelo. *Keketso ya theko ya petrolo e tla ba le sekgahla se seholo thekong ya dijo.*

Eng: (n) **source** the place from which something comes. *Oranges are a very good source of vitamin C.*

Ses: **mohlodi** sebaka seo ntho e itseng e tswang ho sona. *Dilamunu ke sesosa se lokileng sa vitamin C.*

In al drie bogenoemde voorbeelde het die kwaliteitskontroleerder aangedui dat die gedefinieerde terme nie in die gebruiksvoorbeelde weergegee word nie. In die eerste gebruiksvoorbeeld is die term *pepeswa* eerder as *totobatsa* gebruik. In die tweede gebruiksvoorbeeld is *e kgolo* met *(se)seholo* vervang. In die derde

gebruiksvoorbeeld moet *sesosa se lokileng sa* met *mohlodi o lokileng wa* vervang word, sodat die gedefinieerde term in die gebruiksvoorbeeld verskyn.

Net soos in die geval van die vertaling van kollokatiewe inligting, moet die ekwivalent van ’n bepaalde gebruiksvoorbeeld in die korpus van die teikentaal geïdentifiseer word, maar daar is reeds aangedui dat hierdie nie ’n haalbare oplossing binne die Suid-Afrikaanse konteks is nie. Die terminoloog, in samewerking met die vertaler en vakspesialis, moet dus terugkeer na studeerkamervoorbeelde. Hoewel dit nie noodwendig ’n werklikheidsgetroue voorbeeld van taalgebruik sal wees nie (sien paragraaf 1.5), sal veeltalige Suid-Afrikaanse studente toegerus wees met gebruiksvoorbeelde.

4.13 Samevatting

Dit is moontlik om gebruiksvoorbeelde wat natuurlik en kenmerkend, informatief en verstaanbaar is vanuit doelgemaakte korpora te onttrek. Die tekort aan Suid-Afrikaanse vakwoordeboeke wat gebruiksvoorbeelde bevat, skep die probleem dat Suid-Afrikaanse studente nie die nodige terminologiese en/of leksikografiese ondersteuning ontvang nie. Deur gebruiksvoorbeelde in die vorm van volledige sinne aan te bied wat tussen 10-25 woorde is met kollokatiewe inligting wat implisiet of eksplisiet aangebied word, word die behoeftes rondom gebruiksvoorbeelde binne die Suid-Afrikaanse konteks aangespreek. Binne hierdie konteks is sowel empiriese, as fraseologiese kollokasies van groot waarde vir die teikengebruiker. Deur sodanige kollokatiewe inligting semi-outomaties te onttrek met behulp van *Word Sketch*, word die kollokatiewe data opgesom en kan die terminoloog beter rekenskap gee van aspekte soos frekwensie en verspreiding van kollokatiewe inligting. Met die profiel van die teikengebruiker in ag genome, asook die funksie van die veeltalige aanlyn termbank, sal dit meer voordelig wees om kollokatiewe inligting eksplisiet, eerder as implisiet, aan te bied. Die GDEX-funksie skei outomaties goeie gebruiksvoorbeelde van swak gebruiksvoorbeelde op grond van voorafgestelde kriteria waaruit die terminoloog die mees gepaste gebruiksvoorbeeld vir die teikengebruiker kan selekteer wat tot ’n mindere of meerdere mate aangepas moet word.

HOOFSTUK 5

Gevolgtrekking

Die doel van hierdie studie was die ontwerp van ’n model vir die daarstel van ’n veeltalige aanlyn termbank vir Suid-Afrikaanse studente om veeltaligheid te bevorder en studente akademiese ondersteuning te bied in hul sterkste taal in verskeie vakgebiede. Die daarstel van so ’n aanlyn termbank was gegrond op stewige teoretiese beginsels van terminologie. Besondere aandag is gegee aan semi-outomatiese termidentifisering en -onttrekking en die verskaffing van gebruiksvoorbeelde en kollokatiewe inligting in ’n aanlyn termbank.

Dit het behels dat daar in hoofstuk 2 ondersoek ingestel is na ’n toereikende teorie van terminologie, d.i. die Kommunikatiewe Teorie van Terminologie (CTT) wat die meerdimensionaliteit van ’n terminologiese eenheid erken. In teenstelling met die Algemene Teorie van Terminologie (GTT), maak die CTT voorsiening dat ’n terminologiese eenheid vanuit drie afsonderlike perspektiewe benader word, aangesien dit onmoontlik is om die vele multidimensionele fasette van ’n terminologiese eenheid op een slag te benader (Cabré 2003:193). Die CTT maak voorsiening dat vakspesifieke kommunikasie binne die betrokke studie doeltreffend plaasvind tussen spesialiste en semi-spesialiste deur middel van vakspesifieke glossariums. Doelgemaakte korpora maak ’n essensiële deel uit van die CTT. Aangesien die gebruik van hierdie tipe korpora binne die Suid-Afrikaanse konteks nie algemene praktyk is nie, moet hulpbronne beskikbaar gestel word sodat terminoloë opleiding kan ontvang in die samestelling en gebruik van doelgemaakte korpora.

In hoofstuk 3 en 4 is sagteware vir semi-outomatiese termonttrekking op verskeie maniere en uit verskillende uitgangspunte toegepas en geëvalueer. Heylen en De Hertog (2015:214) meld: “Despite the large body of research, there is no generally agreed standard of what a good automatic term extractor should achieve.” Die resultate wat sagteware vir semi-outomatiese termonttrekking oplewer, word eerstens bepaal deur die doel waarvoor dit gebruik word, byvoorbeeld termidentifisering of die onttrekking van gebruiksvoorbeelde. Tweedens word die resultate ook deur die doelgemaakte korpus beïnvloed,

asook die brontaal van so ’n korpus. Indien die resultate nie na wense is nie, moet dit in gedagte gehou word dat daar nie noodwendig iets skort met die sagteware vir semi-outomatiese onttrekking nie – die moontlikheid dat die omvang van die doelgemaakte korpus nie voldoende is nie, moet ten minste oorweeg word (Frankenberg-Garcia 2012:289). ’n Moontlike oplossing is dat die terminoloog meer as een doelgemaakte korpus moet raadpleeg om die gewenste resultate te verkry. Die mens sal altyd die finale seggenskap hê oor die resultate wat hierdie tipe sagteware oplewer – dit moet deeglik deur die terminoloog én vakspesialis bestudeer word. Om die volle omvang van die potensiële resultate wat deur *WordSmith Tools* en *Sketch Engine* opgelewer kan word, te bepaal, word ’n omvattende ondersoek vereis van dié sagteware. Ondersoek moet ook ingestel word na *TickBox Lexicography* (TBL) wat gebruiksvoorbeelde onttrek deur *Word Sketch* en *Good Dictionary Example eXtractor* (GDEX) te kombineer. *TickBox Lexicography* het die potensiaal om belowende resultate op te lewer op ’n koste- en tydeffektiewe manier.

Binne die Suid-Afrikaanse konteks word weinig of geen kollokatiewe inligting in vakspesifieke woordeboeke of aanlyn termbanke weergegee nie. Om aan die behoeftes van die veronderstelde teikengebruiker te voldoen, is hierdie tipe inligting onontbeerlik. Taljard (2016:559) beweer in dié verband: “The advantages of access to LSP corpora, new developments in e-lexicography and the availability of sophisticated software make this a feasible undertaking.”

Aangesien kollokatiewe inligting en gebruiksvoorbeelde nie net bloot vertaal kan word nie en daar ’n tekort aan vakspesifieke tekste in die Afrikatale is, is dit belangrik dat Suid-Afrikaanse universiteite en/of akademiese instansies hulpbronne beskikbaar stel, sodat werkswinkels aangebied kan word waar terminoloë, in samewerking met vertalers en vakspesialiste, ekwivalente van kollokatiewe inligting en gebruiksvoorbeelde in die teikentale kan skep.

Dit is ook belangrik dat nuutskeppings met alle Suid-Afrikaanse universiteite en/of akademiese instansies gedeel word, sodat die Afrikatale bevorder word. Deur dié data binne ’n aanlyn termbank beskikbaar te stel wat toeganklik is vir alle Suid-Afrikaanse studente, is dit haalbaar om die Afrikatale op hierdie manier te bevorder.

So ’n veeltalige aanlyn termbank sal egter net suksesvol wees, indien daar eerstens samewerking tussen die verskillende Suid-Afrikaanse universiteite en/of akademiese instansies is wat hulpbronne beskikbaar stel. Daar is tans talle inisiatiewe by verskillende Suid-Afrikaanse universiteite om terminologie vir spesifiek die Afrikatale beskikbaar te stel. Die Universiteit van KwaZulu-Natal (UKZN) stel byvoorbeeld vakspesifieke woordeboeke saam wat belyn is met dié universiteit se taalbeleid en -plan (Khumalo 2015:495). Hulle het tot op datum 10 vakspesifieke woordeboeke saamgestel wat in Engels en isiZulu beskikbaar is (https://ukzntermbank.ukzn.ac.za/PublicSearch.aspx). Die Universiteit van Kaapstad (UK) het op hul beurt deur hul Veeltaligheidsonderwysprojek 3 vakspesifieke glossariums in die 11 amptelike tale saamgestel (www.mep.uct.ac.za/mep/proj/multi). Die Universiteit Stellenbosch (US) se inisiatief, die selfoonwoordeboek *MobiLex* (www0.sun.ac.za/mobilex/), gee hul studente toegang tot ses vakspesifieke woordeboeke in Engels, Afrikaans en isiXhosa (www.sun.ac.za). Dié universiteit bied ook aan die publiek gratis toegang tot twee vakspesifieke glossariums.

Dit moet beklemtoon word dat die Afrikatale arm is aan elektroniese terminologiese hulpbronne, daarom is dit belangrik dat bestaande data gedeel word. Daar bestaan egter ’n onwilligheid / versigtigheid onder Suid-Afrikaanse universiteite en/of akademiese instansies om bestaande data te deel. Hierdie onwilligheid kan moontlik toegeskryf word aan oorwegings rakende kopiereg en intellektuele eiendom. Daar bestaan heelwat onkunde en onsekerheid by universiteitsowerhede oor die presiese aard van die data wat benodig word, en ook die doel waarvoor dit aangewend word. ’n Aanlyn termbank kan so ontwerp word dat die bron van ’n betrokke termlys die nodige erkenning kry. Samestellers van termlyste kan ook besluit onder watter *Creative Commons* lisensie hulle hul werk wil publiseer; die samesteller kan vanuit ses lisensies kies: (1) *Attribution CC BY*, (2) *Attribution-ShareAlike CC BY-SA*, (3) *Attribution-NoDerivs CC BY-ND*, (4) *Attribution-NonCommercial CC BY-NC*, (5) *Attribution-NonCommercial-ShareAlike CC BY-NC-SA* en (6) *Attribution-NonCommercial-NoDerivs CC BY-NC-ND*. Elkeen van hierdie tipe lisensies word vervolgens kortliks bespreek (https://creativecommons.org/licenses/).

Die *Attribution CC BY* lisensie stel mense in staat om die lisensiehouer se werk op ’n kommersiële vlak te versprei, die volgorde te herrangskik, aan te pas en op die bestaande werk te bou, mits die lisensiehouer erkenning kry vir die oorspronklike samestelling van die data. Die *Attribution-ShareAlike CC BY-SA* lisensie, wat ook vergelyk word met *copyleft* en *open source* sagteware lisensies wat vir kommersiële doeleindes gebruik word, maak voorsiening dat die lisensiehouer se data se volgorde geherrangskik en aangepas word, asook dat daar op die bestaande data gebou word. Alle nuwe data moet onder dieselfde lisensie as die oorspronklike werk gedek word. Onder die *Attribution-NoDerivs CC BY-ND* lisensie word daar voorsiening gemaak vir herdistrubisie op ’n kommersiële en niekommersiële vlak. Die data mag nie aangepas word nie en volle erkenning word aan die lisensiehouer gegee. Indien daar op die *Attribution-NonCommercial CC BY-NC* lisensie besluit word, kan die data se volgorde geherrangskik en aangepas word en mag daar op die bestaande data gebou word, indien dit vir niekommersiële doeleindes gebruik gaan word. Die lisensie vereis dat die lisensiehouer nog steeds onder die nuwe data erken word en die nuwe data moet ook vir niekommersiële doeleindes gebruik word. Die nuwe data hoef egter nie volgens die betrokke vereistes gelisensieer word nie. Sou ’n lisensiehouer op die *Attribution-NonCommercial-ShareAlike CC BY-NC-SA* lisensie besluit, kan die data geherrangskik en aangepas word en daar mag op die bestaande data gebou word wat vir niekommersiële doeleindes gebruik sal word. Die lisensiehouer ontvang erkenning en die nuwe data moet volgens dieselfde voorwaardes gelisensieer word. Die *Attribution-NonCommercial-NoDerivs CC BY-NC-ND* lisensie het die meeste beperkinge van al die *Creative Commons* lisensies. Die lisensiehouer se data mag afgelaai en met ander gedeel word en hy/sy moet nog steeds erkenning daarvoor ontvang. Die data kan nie geherrangskik of aangepas word nie en mag ook nie vir kommersiële doeleindes gebruik word nie. Dit is dus duidelik dat daar wel meganismes bestaan om data behoorlik te lisensieer sodat (a) die nodige erkenning aan die verskaffer van die data gegee kan word, en (b) intellektuele eiendom en kopiereg na behore gereguleer word.

Die studie het verder bevestig dat die betrokkenheid van vakspesialiste noodsaaklik is om die akkuraatheid van die data te verseker. Dit sal voordelig

wees indien twee of meer vakspesialiste per vakgebied as kwaliteitskontroleerders betrokke kan wees. Laastens sal die terugvoer van teikengebruikers verseker dat daar in hul behoeftes voldoen word. ’n Toegekende skakel by elke inskrywing sal die teikengebruiker in staat stel om terugvoering oor ’n spesifieke term, definisie, gebruiksvoorbeeld én kollokatiewe inligting te gee.

BIBLIOGRAFIE

Alberts, M. 2017. *Terminology and terminography principles and practice: A South African perspective*. Cape Town: McGillivray Linnegar Associates.

About the licenses, besigtig op 19 Januarie 2018, beskikbaar by: https://creativecommons.org/licenses/.

Atkins, B.T.S. en Rundell, M. 2008. *The Oxford Guide to Practical Lexicography.* 1st edition. New York: Oxford University Press.

Baker, P. 2004. Querying *KeyWord*s: Questions of Difference, Frequency, and Sense in *KeyWord*s Analysis. In *Journal of English Linguistics, Vol. 32 / No.4*, pp. 346-359.

BootCaT, sagteware vir semi-outomatiese termonttrekking, besigtig op 26 Mei 2017, beskikbaar by: http://bootcat.dipintra.it/

Bowker, L. en Pearson, J. 2002. *Writing with Specialized Corpora: A practical guide to using corpora.* London: Routledge.

Bureau for Institutional Research & Planning, University of Pretoria.

Cabré, M.T. Theories of Terminology. Their description, prescription and explanation. In *Terminology*. 2003, pp. 163–199.

Cabré, M.T. Montané, M.A. & Nazar, R. 2012. *Corpus-based terminology* processing, besigtig op 6 Junie 2016, beskikbaar by: http://terminus.iula.upf.edu/tke2012/.

Castagnoli, S. Using the Web as a Source of LSP Corpora in the Terminology Classroom. In *Marco Baroni and Silvia Bernardini (eds.) Wacky! Working papers on the Web as Corpus. Bologna: GEDIT*. 2006, pp. 159-172.

Coetzee-Van Rooy, A.S. 2010. *The importance of being multilingual*. Inaugural lecture, North West University Vaal Triangle Campus.

Concepts of profit and loss and other comprehensive income, besigtig op 21 Januarie 2018, beskikbaar by: www.accaglobal.com.

Costa, H. Pastor, G.C. en Seghiri, M. 2016. Nine Terminology Extraction Tools: Are they useful for translators? In *Multilingual, Vol. 27, No. 3,* pp. 14-20.

DANTE, databasis, besigtig op 23 November 2017, beskikbaar by: http://www.webdante.com/the_database.html.

Evert, S. 2007, *Corpora and collocations*, besigtig op 23 November 2017, beskikbaar by: http://www.stefan-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf.

Faber Benítez, F. 2009. The Cognitive Shift in Terminology and Specialized Translation. In *MonTI, núm. 1*, pp.107-134.

Feurtes-Olivera, P.A. en Arribas-Baño, A. 2008. *Pedagogical Specialised Lexicography: The representation of meaning in English and Spanish business dictionaries.* Amsterdam: John Benjamins Publishing Co.

Frankenberg-Garcia, A. 2012. Learners' Use of Corpus Examples. In *International Journal of Lexicography,* vol. 25, no.3*,* pp.273-296.

Heylen, K. & De Hertog, D. 2015. Automatic Term Extraction. In *Handbook of Terminology, vol. 1*, pp. 199-219.

Hiles, L. 2009. *Examples in South African School Dictionaries: from Theory to Practice*. Master of Philosophy, Stellenbosch University.

Humblé, P. 2001. *Dictionaries and Language Learners*. Frankfurt: Haag & Herchen.

Joffe, D. & de Schryver, G.-M. TshwaneLex A State-of-the-Art Dictionary Compilation Program. In *Proceedings of the 11th EURALEX International Congress*, pp. 99-104. Lorient, France.

Khumalo, L. 2015. Semi-automatic Term Extraction for an isiZulu Linguistic Terms Dictionary*. In *Lexikos 25*, pp. 495-506.

Kilgarriff A. & Rundell, M. 2002. Lexical Profiling Software and its lexicographic applications: a case study. In *Proceedings of EURALEX International Congress*, pp. 807-818. Copenhagen.

Kilgarriff, A. Rychly, P. Smrz, P. en Tugwell, D. 2004. The *Sketch Engine*. In *Proceedings of the 11th EURALEX International Congress* pp. 105-116. Lorient, France.

Kilgarriff, A. Miloš, H. Mcadam, K. Rundell, M. en Rychlý, P. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the 13th EURALEX International Congress.* pp. 425-432.

Kilgarriff, A. Kovář, V. & Rychlý, P. 2010. Tickbox Lexicography. In *eLexicography in the 21st century: New challenges, new applications.* pp. 411-418. Presses universitaires de Louvain, Brussels.

Kilgarriff, A. Vojtech, Krek, S. Srdanović, I & Tiberius, C. 2010. A quantitative evaluation of word sketches. In *Proceedings of the 14th EURALEX International Congress*, pp. 372-379. Leeuwarden, the Netherlands.

Kilgarriff, A. & Kosem, I. 2012. Corpus tools for lexicographers. In *Electronic Lexicography, edited by Granger, S. & Paquot, M.* pp. 31-56. Oxford University Press, United Kingdom.

Kilgarriff, A. Baisa, V. Bušta, J. Jakubíček, M. Kovář, V. Michelfeit, J. Rychlý, P. & Suchomel, V. 2014. The *Sketch Engine*: Ten Years On. In *Lexicography 1(1)*, pp. 7-36.

Kosem, I. 2016. Interrogating a corpus. In *The Oxford Handbook of Lexicography, edited by Philip Durkin*. pp. 76-93. Oxford University Press, United Kingdom.

Kübler, N. & Pecman, M. 2012. The ARTES bilingual LSP dictionary. In *Electronic Lexicography, edited by Granger, S. & Paquot, M.* pp. 187-209. Oxford University Press, United Kingdom.

L'Homme, M.-C. Robichaud, B. & Leroyer, P. 2012. Encoding collocations in the DiCoInfo. In *Electronic Lexicography, edited by Granger, S. & Paquot, M.* pp. 211-236. Oxford University Press, United Kingdom.

Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography*. 2nd edition. Cambridge: Cambridge University Press.

Lin, D. 1998. Extracting collocations from text corpora. In *Proceedings of the First Workshop on Computational Terminology*, pp. 57-63. Montreal, Canada.

McEnery, T. Xiao, R. & Tono, Y. 2006. *Corpus-Bases Language Studies: An advanced resource book*. London and New York: Routledge Taylor & Francis Group.

McKeown, K. & Dragomir, R.R. 2000. Collocations. In *Handbook of Natural Language Processing, Dekker, M*.

MobiLex, selfoonwoordeboek, besigtig op 19 Januarie 2017, beskikbaar by: www0.sun.ac.za/mobilex/.

Muegge, U. *10 things you should know about automatic terminology extraction* 2013, besigtig op 23 Oktober 2017, beskikbaar by: https://www.linguagreca.com.

Multilingual Glossaries Project, besigtig op 19 Januarie 2018, beskibaar by: www.mep.uct.ac.za/mep/proj/multi.

MULTILINGUAL MOBILE DICTIONARY HELP STUDENTS MASTER CONCEPTS. Besigtig op 19 Januarie 2018, beskikbaar by: www0.sun.ac.za.

National Benchmark Tests Project & standards for National Examination & Assessment Systems: Department of Higher Education, soos op 18 Augustus 2009, besigtig op 18 Mei 2016, beskikbaar by: https://pmg.org.za/committee-meeting/10668/.

Nchabeleng, M.J. 2011. *Terminological Issues in the Translation of Chemistry Terms from English to Northern Sotho*. Mini-dissertation, University of Pretoria.

Nesselhauf, N. 2003. The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. In *Applied Linguistics 24/2,* pp. 223-242.

Otto, A. 1998. Die onderrig van kollokasies aan niemoedertaalsprekers binne die Suid-Afrikaanse tersiêre konteks. In *South African Journal of Linguistics 16:sup35,* pp. 99-112.

OERTB, aanlyn termbank, besigtig op 21 September 2017, beskikbaar by: oertb.tlterm.com.

Okapi Framework, sagteware vir semi-outomatiese termonttrekking, besigtig op 19 Januarie 2018, beskikbaar by: okapi.sourceforge.net.

Packheiser, K. 2009. *The General Theory of Terminology: A Literature Review and Critical Discussion*. Master Thesis, Copenhagen Business School.

Pearson, J. 1998. *Terms in Context.* Amsterdam: John Benjamins Publishing Company.

Reiplinger, M. Schäfer, U. en Wolska, M. 2012. Extracting glossary sentences from scholarly articles: A comparative evaluation of pattern bootstrapping and deep analysis. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries,* pp.55-65.

Rundell, M. 1998. Recent Trends in English Pedagogical Lexicography. International Journal of Lexicography, vol. 11, no. 4, pp.315-342.

Rundell, M. 2012. The road to automated lexicography: An editor's viewpoint. In *Electronic Lexicography, edited by Granger, S. & Paquot, M.* pp.15-29. Oxford University Press, United Kingdom.

Sageder, D. Terminology Today: A Science, an Art or a Practice? Some Aspects of Terminology and its Development. In *Brno Studies in English 36(1)*. 2010, pp. 123-134.

Scott, M. 1997. PC ANALYSIS OF KEY WORDS – AND KEY KEY WORDS. In *System 25(1)*, pp. 233-245.

Simple Extractor, sagteware vir semi-outomatiese termonttrekking, besigtig op 19 Januarie 2018, beskikbaar by: www.dail-software.com.

Sketch Engine, sagteware vir semi-outomatiese termonttrekking, besigtig op 19 Desember 2017, beskikbaar by: www.sketchengine.co.uk.

Statistics used in the Sketch Engine, 2015, besigtig op 05 November 2017, beskikbaar by: https://www.sketchengine.co.uk/wp-content/uploads/ske-statistics.pdf.

Stubbs, M. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. In *Functions of Language, 1*. pp. 23–55.

Srdanović, I. Ida, N. Shigemori Bućar, C. Kilgarriff, A. & Kovář, V. 2011. Japanese word sketches: Advantages and problems. In *Acta Linguistica Asiatica* 1(2), pp. 63-82.

Taljard, E. & De Schryver, M.-G. 2002. Semi-automatic Term Extraction for the African Languages, with Special Reference to Northern Sotho. In *Lexikos 12*, pp. 44-74.

Taljard, E. 2004. Semi-Automatic Retrieval of Definitional Information: A Northern Sotho Case Study. In *Leksikos 14*, pp.173-194.

Taljard, E. 2015. Collocations and Grammatical Patterns in a Multilingual Online Term Bank. In *Lexikos 25*, pp. 387-402.

Taljard, E. 2016. Collocational information for terminological purposes. In *Proceedings of 17th EURALEX International Congress,* pp. 553-560. Tbilisi, Georgia.

Temmerman, R. 2000. *Towards New Ways of Terminology Description: The sociocognitive approach*. John Benjamins Publishing Company.

Terminus, sagteware vir semi-outomatiese termonttrekking, besigtig op 19 Januarie 2018, beskikbaar by: terminus.iula.upf.adu.

Trask, R.L. 1993. *A Dictionary of Grammatical Terms in Linguistics.* Psychology Press.

UKZN termbank, aanlyn termbank, besigtig op 24 Januarie 2018, beskikbaar by: https://ukzntermbank.ukzn.ac.za/PublicSearch.aspx.

VAT, besigtig op 20 Januarie 2018, beskikbaar by: www.sars.gov.za.

Warburton, K. 2014. Glossary Creation Service leveraging: state-of-the-art Term Extraction. In*Termologic*, pp.1-6.

WordSmith Tools, sagteware vir semi-outomatiese termonttrekking, besigtig op 26 Oktober 2017, beskikbaar by: www.lexically.net/wordsmith/.