# HET GEBRUIK VAN CONCORDANTIES EN INDICES BIJ ONDERZOEKINGEN OP HET GEBIED VAN HET ZESTIENDE-EEUWSE EN HET MODERNE NEDERLANDS

S. DE VRIENDT
Professor aan de Ecole d'interprètes internationaux
docent aan de V.U.B. - U.L.B.

Deze bijdrage zou aanvankelijk een verslag zijn van de projecten en verwezenlijkingen van de Fakulteit der Letteren en Wijsbegeerte van de V.U.B. Doordat echter

1. de twee afgewerkte projecten het werk zijn van docenten die tevens verbonden zijn aan de ULB (Sectie Germaanse filologie van de Fakulteit der Letteren en Wijsbegeerte),
2. voor die twee projecten een deel van het computerwerk verricht werd op het rekencentrum van het Centre universitaire de l'Etat te Bergen (Mons) en in samenwerking met leden van dit rekencentrum,
3. schrijver dezes ook professor is aan het Centre Universitaire de l'Etat te Bergen, en er aan andere projecten deelgenomen heeft, is dit uitgegroeid tot een rapport over verwezenlijkingen en plannen van de germanisten van VUB en ULB en over het werk dat verricht werd en wordt op de tolkenschool (E.I.I.) en het Département de Linguistique van het universitair centrum te Bergen.

Gezien de beperkte ruimte waarover ik beschik is het uitgesloten dat van elk der projecten een uitvoerige beschrijving zou gegeven worden. Ik zal dit alleen doen voor de eerste drie van de in de volgende lijst opgesomde projecten, met tussen haakjes de vermelding van de naam van de leider(s) :

1. De taal van twee Vlaamse handschriften uit de tweede helft van de 16de eeuw (S. de Vriendt, Brussel en Mons);
2. Frequentie van woorden en structuren in spontaan gesproken Nederlands (M.J. de Vriendt-de Man, Brussel en Mons);
3. Hooft : Granida (S. de Vriendt en de Sectie Nederlands, Mons);
4. Woordsoorten in het Frans en het Nederlands (A. Beerden, Mons) - Geponst worden niet-gedialogeerde delen uit Franse romans en hun vertaling in het Nederlands evenals delen uit Nederlandse romans en de Franse vertaling hiervan;
5. Franse woorden in Nederlandse kranten (M.J. de Vriendt-de Man, licentieverhandeling van J. Dupont, Mons). Het corpus bestaat uit artikelen over film, muziek en beeldende kunsten;
6. Duitse voorzetsels en hun vertaling in het Frans (S. de Vriendt, licentieverhandeling van J. Tamigniaux, Mons). Het corpus bestaat uit de eerste honderd bladzijden van de Erinnerungen 1945-1953 van C. Adenauer en de vertaling hiervan (Mémoires) door Denise Meunier;
7. Simulatie van het aanleren van een taal door middel van een computer (J.J. Waaub, Mons). Het doel is het testen van de progressiviteit van een cursus, in casu de audiovisuele cursus Engels van Filipovic en Webster;



Extrait de la Revue (R.E.L.O.)
VII, 1 à 4, 1971. C.I.P.L. - Université de Liège - Tous droits réservés.

8. Gesubstitueerde en substituerende woorden in Tardieus *Un Mot pour un autre* (J. de Decker, Mons);
9. Woordenschat en grammatische structuren in een Russisch scheikundig handboek (C. Carboni, Mons).
   Het doel van die projecten is niet zozeer het samenstellen van een microglossarium van scheikundige termen maar wel het bepalen van de grammatische structuren en de niet specifiek scheikundige woordenschat die in een dergelijk handboek wél en in audiovisuele basiscursus, waarmee de studenten de taal leren, niet voorkomen.

De projekten 1, 2, 5 en 8 zijn beëindigd.

*De taal van twee Vlaamse handschriften uit de tweede helft van de zestiende eeuw.*

Die twee handschriften zijn nog onuitgegeven reisjournaals die in het vervolg met "journaal van Karel Rijm" en "journaal van Levinus Rijm" zullen aangeduid worden. Karel Rijm was een Gentenaar die in 1570 door de Oostenrijkse Keizer als Ambassadeur naar Constantinopel werd gestuurd, waar zijn voorganger, Albert de Wijs, zelf de opvolger van de beroemde Ogier Gisleen van Busbeek, in 1569 overleden was. In zijn journaal vinden we een beschrijving van de heenreis, van Brussel naar Wenen en Praag en dan verder naar Constantinopel evenals, maar dan in veel beknoptere vorm, van de terugreis vier jaar later. Minder dan tien jaar later vertrok zijn broer Levinus eveneens naar de Turkse hoofdstad, in het gevolg van Paul Freiherr von Eytzing. Dit tweede journaal is wat de inhoud betreft interessant omdat het ook iets vertelt over het verblijf in Turkije : voor de filoloog is echter van groter belang dat de schrijver van het tweede journaal op talrijke plaatsen het eerste overgeschreven heeft, met tal van varianten vooral op het gebied van de spelling en de morfologie. Die beide teksten heb ik getranscribeerd en daarna direct zelf geponst (1). Elke kaart kreeg een codering (tekst, bladzijde, kaartnummering). De afzonderlijke woorden werden niet gecodeerd : speciale tekens werden gebruikt om de doorgehaalde tekstgedeelten aan te geven. Op elke tekst heb ik een index en een concordantie laten maken. Een index is, zoals men weet, "een alfabetische lijst van alle woorden die in een tekst voorkomen, met vermelding van hun vindplaatsen, doch zonder enige context" (F.J.M.L. de Tollenaere in "Taal-onderzoek" in onze Tijd, Den Haag, 1962, p. 94).

Het grootste voordeel van de index is dat hij deze lijst in een vrij geconcentreerde vorm levert. Het reisjournaal van K. Rijm b.v. bevat 11.196 woorden (N = 11.196, dus 11.196 *tokens*) maar slechts 2.473 verschillende woorden (V = 2.473, dus 2.473 *types*). Een lijst van die 2.473 types, met vermelding van de vindplaatsen, zou, op twee kolommen, ongeveer vijftig bladzijden vullen. Wie de index gebruikt kan zonder veel moeite de plaats(en) in de tekst terugvinden waar het woord voorkomt, zonder veel moeite althans als het gaat om woorden met een vrij lage frequentie. Maar wat moeten we aanvangen als we zien dat er tientallen of honderden voorbeelden zijn van dat woord : in mijn eerste handschrift 186 x *in*, 242 x *ende*, enz. In het onderzoek naar de woordfrequentie in kranten van de stichting Mathematisch Centrum zijn er 1083 voorbeelden van *in* en 1852 van het type *van* !

De conclusie ligt voor de hand : de index is een uitstekend hulpmiddel voor de lexicoloog. Hij vindt daar

heel vlug de woorden die hem interesseren en dat zijn inderdaad vaak woorden met een vrij gering aantal vindplaatsen. Voor de grammaticus is het gebruik van een index, beter gezegd van uitsluitend een index van weinig nut. Een concordantie is een index die voor elk woord, elke vorm van een woord, "vergezeld is van een zekere portie context" (De Tollenaere, op.cit. p. 99). Elk woord betekent hier dus elk token. De opgenomen portie context kan zo kort en zo lang zijn als men wil. Men mag echter niet vergeten dat hoe groter de context, hoe omvangrijker de concordantie wordt. Voor mijn handschriften was voor elk token de portie tekst beperkt tot 132 plaatsen (letters, leestekens en spaties, d.i. één regel) met het keyword in het midden. Mijn concordantie op het journaal van K. Rijm telt dus 11.196 regels, ongeveer 375 grote bladen.

Het is duidelijk dat een concordantie op langere teksten al te lijvig en daardoor onbruikbaar wordt. "Geen wonder dan ook", schrijft De Tollenaere (dp.cit. p. 99 en 101), "dat bij concordanties meestal de grammaticaal-technische woorden (Fr. *mots de structure* of *mots-outils*), d.w.z. lidwoorden, voorzetsels, hulpwerkwoorden enz., worden weggelaten, wat neerkomt op niets minder dan de helft van het totale aantal woorden" "in een concordantie heeft men alleen maar literaire of theologische, en stellig geen woordstatistische doeleinden op het oog".

Maar als men die wél heeft, en nog meer als men een kwantitatief onderzoek naar grammatikale vormen en structuren op het oog heeft, dan mogen die grammaticaal-technische woorden niet weggelaten worden. Doordat de teksten die ik wou bestuderen vrij kort waren bood de omvang van de concordanties voor mij geen enkele moeilijkheid. Bovendien maakte het feit, dat bij elk token in de concordantie ook de vindplaats vermeld was, de lijst der vindplaatsen in de index overbodig. Ik liet deze dus eenvoudig optellen zodat de index alleen het aantal vindplaatsen vermeldt.

Concordantie en index bleken elkaar voor het geplande onderzoek uitstekend aan te vullen. Immers de concordantie stelde me in staat

1. de in de index voorkomende vormen correct te interpreteren :
   *acht* = 8, niet (ik) acht
   *achten* = 8, niet het ww. achten
   *berste* = beste, niet van het ww. bersten, enz.

2. de in de index voorkomende homografen te scheiden
   bv. *haudde* = oude
   *haudden* = houden
   *haude* (11x) = oude
   *hauden* (5x) = 4x houden, 1x ouden
   *haut* (8x) = 1x houdt, 7x oud
   *haute (2x)* = hout
   *hauten (2x)* = houten
   *hauts* = (van) ouds

3. de context van de woorden te bestuderen.

Aan de hand van index en concordantie heb ik toen de taal van de beide handschriften beschreven. Deze studie die ook de tekst van de beide handschriften, en de indices bevat - met later toegevoegde vermelding der vindplaatsen voor sommige woorden en een paar woordwerklaringen - wordt dit jaar uitgegeven door de Koninklijke Vlaamse Academie voor Taal- en Letterkunde. Ik beperk me dus verder tot kommentaar over

a. woordcodering
b. een groter corpus, andere mogelijke vormen van automatisering.

a. Voor het hier besproken onderzoek heb ik afgezien van elke woordcodering. Dit betekent geenszins dat ik van mening ben dat een dergelijke codering steeds uit den boze zou zijn (cf. het Granida-project). Het is echter wel moeilijk een bevredigend coderingssysteem te vinden. Voor het onderzoek van de Stichting Mathematisch Centrum over 50.000 woorden uit kranteartikelen heeft men zich beperkt tot vier categorieën, nl.

   werkwoorden 3
   bijvoeglijke woorden (bn en bw) 2
   zelfstandige naamwoorden 1
   de rest 0

   Ik heb de indruk, maar ik kan me vergissen, dat zo'n klein aantal categorieën van weinig nut is. Daarbij is het niet altijd gemakkelijk te bepalen tot welke groep elk woord behoort. Geeft men de voorkeur aan een groter aantal categorieën, dan wordt het moeilijk dit aantal te beperken, mede omdat het aantal grensgevallen eerder toe - dan afneemt. Vooral bij oudere teksten vind ik dat het coderen van de afzonderlijke woorden moeilijk is, zoals b.v. *alsser* (als er), *aleyst dat* dat op andere plaatsen *al eyst dat* (al is't dat, = al) geschreven wordt. Hoe zal men *hoewel* coderen dat soms aan elkaar, soms in twee woorden (*hoe wel*) geschreven wordt ? Zelfs als men meent voor elk dezer gevallen een goede oplossing te kunnen vinden, dan moeten die en gelijkaardige gevallen, in het hele corpus, altijd op dezelfde manier behandeld worden. Verder ben ik bang dat de filoloog zijn kennis van het huidige Nederlands en van andere teksten uit vroegere eeuwen in de zinsontleding zal projecteren, die het coderen van de woorden min of meer bewust voorafgaat.
   Zo leest men b.v. in het journaal van K. Rijm, dat nagenoeg geen leestekens vertoont :

   "zijn ghereyst dien dach tot Dicermunt *dat es* vier myllen van Weenen in Oostenryck".

   Als men dit leest met een pauze na Dicermunt en dus *dat* als een aanwijzend voornaamwoord beschouwt en het ook als dusdanig codeert en als men hetzelfde doet voor alle zinnen van dit type, zal men licht tot de conclusie komen, dat de persoonsvorm in alle zgn. bijzinnen in het eindstuk van deze constructies staat. Men zou daarbij over het hoofd gezien hebben, dat het woordje *dat* hier ook betrekkelijk voornaamwoord kon zijn en dat de verhouding tussen het aantal bijzinnen met de persoonsvorm in het eindstuk en het aantal bijzinnen die dit kenmerk niet vertonen heel anders zou kunnen zijn dan

men dacht. Maar wie kan in de twintigste eeuw weten hoe een bepaalde zestiende-eeuwer de structuur van een bepaalde zin opvatte. Een laatste bezwaar tegen woordcodering is dat, als iedereen niet dezelfde codering gebruikt, de vergelijking van resultaten van verschillende onderzoekingen een uiterst delicate zaak wordt.

b. Met andere mogelijke vormen van automatisering bedoel ik dat m.i. tal van problemen kunnen worden opgelost als de taalkundige steeds over de mogelijkheid beschikt nieuwe vragen te stellen aan de computer. Dit is het geval als het behandelde corpus zo uitgebreid is dat niet kan gedacht worden aan het maken van een concordantie. Het verstandigste lijkt mij dan te zijn dat men eerst machinaal een index met opgave van vindplaatsen en frequentie van voorkomen laat maken. Wie dan één bepaald probleem wil onderzoeken, kan een gedeeltelijke concordantie laten maken. Een paar voorbeelden :

- in het journaal van L. Rijm lezen we *beddekeens*. Een concordantie op alle woorden die op -eens uitgaan zou een drietal andere diminutiva bevatten (*bergelkeens, hullekeens, husekeens/hwskeens*) maar ook adverbia (*smurgeens, tsaueens, tsmurgeens* enz.), voorzetsels (*behaudeens, neffeens*) en meervouden van substantieven (*guldeens, wageens*). Verder onderzoek zou dan uitwijzen dat deze verdubbeling alleen voor *-ns* optreedt (slechts 1x *conyngheene*).

- de vraag van de taalkundige moet zo gesteld worden dat geen enkele vorm die hij wenst te krijgen ontbreekt, maar ook dat slechts een beperkt aantal ongewenste vormen (*noises*) opgenomen wordt. Voor mijn handschriften zou een concordantie op alle woorden uitgaande op *-nde* b.v. alle onvoltooide deelwoorden bevatten en dus de beschrijving van de structuur van de constructies, waarvan zij de kern zijn, mogelijk maken. Het aantal "noises" zou beperkt zijn tot negen types (*alderande, allederande, ande, ....*) waarvan alleen *vande* ( = van de ) door een vrij groot aantal tokens vertegenwoordigd is.

*Frequentie van woorden en structuren in spontaan gesproken Nederlands.*

In de jaren na de tweede Wereldoorlog heeft in ons land het inzicht veld gewonnen dat de woordenschat die men de leerlingen in het vreemde talenonderwijs zou aanbieden niet alleen kwantitatief maar ook kwalitatief gedoseerd moest worden en dat dit het best kon gedaan worden door de leerlingen eerst de meest frequente woorden te laten leren, daarna de wat minder frequente, enz. Voor de tweede taal in onze franstalige scholen, het Nederlands, beschikte men toen al over de lijst van De la Court. Deze frequentielijst was gebaseerd op geschreven, beter gezegd gedrukt materiaal uit het einde van de 19de en de eerste decennia van de 20ste eeuw. Onder de recentere frequentietellingen is er tot nog toe maar één waarvoor het onderzochte corpus uit de tekst van opgenomen Nederlandse gesprekken bestaat en dat is het project waarvan hier de voornaamste kenmerken geschetst zullen worden.

Het materiaal bestaat uit 29 gesprekken, 16 Noordnederlandse en 13 Zuidnederlandse (twee derde gesprekken die in de woonkamer of in een winkel opgenomen werden, één derde informele, niet voorbereide interviews van de B.R.T. en de Nederlandse Radio Unie), met samen 117.122 woorden (58.578 voor de Noordnederlandse, 58.544 voor de Zuidnederlandse opnamen). Het werd op ponskaarten gebracht, zonder woordcodering, en daarop werd een concordantie gemaakt. De vormen werden niet gelemmatiseerd, de homografen echter wel met de hand gescheiden. Aan de hand van de concordantie werd, eveneens met de hand, de range of spreiding van elke vorm over de 29 gesprekken, bepaald. Het resultaat hiervan was enerzijds een index met, voor elk woord, de frequentie en de range in het Noordnederlandse en het Zuidnederlandse materiaal en anderzijds een frequentielijst per woordsoort.
De verkregen lijsten werden toen vergeleken met die van De la Court en die van Van Berckel (Stichting Mathematisch Centrum).
Verder werden ook nog kwantitatief beschreven de combinaties van persoonsvorm met infinitief en voltooid deelwoord, de plaats van de elementen van deze combinaties aan het einde van de zgn. bijzinnen (met of zonder uitloop), de manier waarop spreekreacties ingezet worden, enz....
Een tweede luik van dit onderzoek werd gevormd door een enquête over de beschikbare woordenschat (1), het "vocabulaire disponible", d.w.z. de woorden die in het gesprek niet heel vaak gebruikt worden en dus in frequentielijsten niet of op de laatste plaatsen te vinden zijn, maar die onontbeerlijk zijn in die zin dat ze onrechtstreeks aanwezig zijn in de taaluitingen, b.v. in de voornaamwoordelijke aanduiding. Zo wordt het woord *bus* niet uitgesproken als men zegt *Wanneer komt ie nu?* of *Daar is ie* maar niemand zal hier *het* gebruiken i.p.v. *ie.*
Dezelfde lijst van 16 onderwerpen, die Gougenheim, Michéa, Rivenc en Sauvageot (zie l'Elaboration du Français Fondamental, Didier, Paris) voor het Frans gebruikt hadden, werd aan schoolgaande kinderen tussen 12 en 16 jaar voorgelegd. Voor elk onderwerp schreven de leerlingen een lijst van twintig zelfstandige naamwoorden op. Het materiaal werd in acht scholen verzameld (vier Nederlandse, vier Belgische, vier jongens- en vier meisjesscholen). De oogst bedroeg wat meer dan 49.000 woorden die door de computer zo gesorteerd werden, dat we nu over een frequentielijst per onderwerp beschikken, maar bovendien ook binnen elk onderwerp, over frequentielijsten voor het Nederlandse en het Belgische materiaal en andere voor de jongens en de meisjes afzonderlijk.

*Hooft : Granida.*

Het derde project is van een heel andere aard. Het uitgangspunt hiervan is te vinden in de werken van Charles Muller over de toneelspelen van P. Corneille. Muller is van mening dat statistische gegevens slechts een bevestiging kunnen brengen van wat de interne studie van de tekst opgeleverd heeft. Ik heb gedacht dat het boeiend kon zijn eens na te gaan of en hoe iemand, die over de statistische gegevens beschikt maar de tekst niet kent, uit deze gegevens informaties kan putten betreffende de inhoud van de tekst. Dit had uiteraard tot gevolg dat één persoon de tekst moest kiezen en ook moest bepalen welke operaties



Extrait de la Revue (R.E.L.O.)
VII, 1 à 4, 1971. C.I.P.L. - Université de Liège - Tous droits réservés.

de tekst zou moeten ondergaan en welke statistische toets of toetsen men zou uitvoeren terwijl één of meer andere personen de statistische gegevens zouden onderzoeken. Het eerste deel van het werk bestond dus in het voorbereiden van het materiaal dat men de proefpersonen wilde voorleggen. De tekst van Hoofts Granida werd geponst met een codering per kaart en per woord. Op elke kaart staat precies een vers, plus (in de eerste kolommen) het nummer van het vers, van het bedrijf en van de scène en ook een codecijfer per personage (er zijn er precies tien). Elke woord kreeg een codecijfer naar gelang van de woordklasse waartoe het behoort. Omdat we de codering tot één cijfer wilden beperken (anders waren sommige verzen te lang geworden voor de beschikbare ruimte), werd de volgende codering toegepast :

0 : Voorzetsels en voegwoorden
1 : werkwoorden
2 : zelfstandige naamwoorden
3 : bijvoeglijke naamwoorden en bijvoeglijk gebruikte voornaamwoorden en telwoorden
4 : bijwoorden
5 : zelfstandig gebruikte voornaamwoorden en telwoorden
6 : eigennamen
7 : scheidbare partikels
8 : samenstellingen van het type *mer* ( = men er), *dats* ( = dat is)
9 : interjecties

De vier proefpersonen, een dokter in de rechten, twee Belgische germanisten en een Nederlandse romanist kregen het volgende materiaal :

1 : indices (zo konden ze weten welk personage een bepaald woord in een bepaalde scène éénmaal of meer dan éénmaal had gebruikt)
2 : een frequentielijst per woordklasse en per personage
3 : een tabel waarvan ze konden aflezen hoeveel woorden van elke woordklasse elk personage in het hele stuk, in elk bedrijf, in elke scène heeft uitgesproken en dus ook welke personages zich in elke scène samen op het toneel bevinden.
4 : de resultaten van een $X^2$- toets waaruit blijkt voor welke woordklassen elk personage een significante afwijking vertoont van een normale distributie, m.a.w. welk personage een tekort heeft aan substantieven of een teveel aan interjecties, enz.

Een belangrijke opmerking is dat de woorden van de zesde klasse (eigennamen) overal meegeteld werden maar uit de indices en frequentielijsten verwijderd werden zodat ze voor de proefpersonen onbekend bleven.
De vier proefpersonen kregen als enige informatie over het stuk dat het geen vertaling was maar een Nederlands stuk uit de zestiende eeuw of zeventiende eeuw. Geen uitleg (*krijcht* b.v. is een vorm v. krijgen = oorlog voeren ! ) werd verstrekt over de woorden in indices en frequentielijst. De proefpersonen werd gevraagd het stuk zoveel mogelijk te reconstrueren. Gesuggereerd werden de volgende vragen : tot welk

genre behoort het stuk ? Welke is de handeling ? Wat zijn de personages ? Wat weet u over hun karakter ? Welke is hun onderlinge verhouding ? De proefpersonen werkten eerst afzonderlijk en schreven een rapport over hun bevindingen. Daarna kwamen ze samen, vergeleken wat ze gevonden hadden en stelden een definitief verslag op.

De voornaamste conclusies die men uit dit onderzoek kan trekken zijn :

1. dat het mogelijk gebleken is belangrijke delen van het stuk te reconstrueren en ook dat het karakter van de personages en de verhoudingen vrij goed "herkend" werden, dit vooral nadat het kleine team de bevindingen had kunnen vergelijken.
2. dat vooral de frequentielijsten nuttig geweest zijn, echter niet uitsluitend doordat deze deden uitkomen welke woorden in elke klasse door een bepaald personage vaak gebruikt waren. Ook de woorden met een lage frequentie (tot en met frequentie 1) waren belangrijk.
3. dat de tabel met het aantal woorden per personage per deel van het stuk nuttig was voor het vaststellen van bepaalde banden tussen personage X en personage Z. De proefpersoon die hieruit iets meende te kunnen afleiden over de aard van die banden ging echter een verkeerde weg op.

Ik hoop dat de beschrijving van deze drie projekten een goed idee geeft van het werk dat door mijn collega's en mijzelf verricht werd en nog wordt. Het bestaande Granida-materiaal wordt verder verwerkt, de projekten 4, 7, 9 zijn aan de gang, andere zijn in voorbereiding, maar het is te vroeg om aan die kinderen al een naam te geven.

(1) Het direct zelf overbrengen op ponskaarten van de tekst van een handschrift is m.i. aanbevelenswaardig. Hierbij gaat immers helemaal geen tijd verloren daar de tekst anders toch overgetikt en eventueel naderhand door een ponstypiste geponst moet worden, die uiteraard met de taal van de tekst veel minder vertrouwd is en daardoor meer fouten kan maken.

m

(2) Voor het eerste deel van dit project werden alle operaties te Brussel verricht (Centre de Linguistique Automatique Appliquee); het materiaal voor de enquete over de beschikbare woordenschat werd te Bergen geponst en gesorteerd.

*SUMMARY*

This paper first presents a short survey of computer-aided studies in the field of philology and linguistics at the Department of Dutch of both Brussels universities and at the Interpreters' School and Department of Linguistics of the Centre Universitaire de l'Etat at Mons. It gives a more detailed account of three projects, the first two of which are now completed :

1) a description of the spelling system, morphology and (partly) syntax of two sixteenth-century Flemish texts : the diaries of an imperial ambassador's journey from Vienna to Constantinople and of a similar journey by the ambassador's brother thirteen years later.

2) a. a frequency-list of words, word-classes and (some) grammatical structures in spoken Dutch;
   b. a study of the "available vocabulary" ("vocabulaire disponible", after Gougenheim, Michéa, Rivenc and Sauvageot) of Dutch and Flemish adolescents.

3) a study of how and to what extent one is able to reconstruct certain features of a play (plot, characters, etc.) from an index, frequency-lists of words, tables indicating how many words of each word-class were spoken by each character in each scene, each act and in the whole play (with $X^2$-test pointing to significant deviations).