Anailís Dioscúrsa ar bonn Téacs

Fionn Murtagh*
Scoil na Ríomhaireachta
Ollscoil na Banríona, Béal Feirste BT7 1NN
fmurtagh@acm.org

Achoimre

Tugaimid súil siar ar shaothar a bhaineann le húsáid ghaol-anailíse ar dhoiciméid agus ar théacsanna. Léirímid buntáistí na hanailíse measta sin ar bonn roinnt ghearr-scéalta de chuid Shéamais Mhic Ghrianna. Ina dhiaidh sin i saothor nuálach leathnaímid an anailís i dtreo scrúdú an dioscúrsa, .i. gluais nó druidim an réasúnú nó na loighce atá léirithe san téacs.

Abstract

"Machine Analysis of Discourse based on Text". We review work related to the use of correspondence analysis on documents and texts. We show the advantages of such quantitative analysis on a number of short stories of Séamas Mac Grianna. Following that, in innovative work, we extend the analysis in the direction of examining discourse, viz. the progression or flow of reasoning or logic represented in the text.

1 Réamhrá

Úsáidtear gaol-anailís (anailís chomhfhreagartha, analyse des correspondances, Benzécri, 1976, 1992) chun dearcadh a fháil ar na gaoil nó na nascanna a féadtar a aimsiú nó a chur in iúl i dtearmann nó i bhfearann faoi leith. Aidhm anailíse mar seo ná léirmhíniú a fháil ar an dtearmann trí nochtadh na ngaol agus na nascanna. Déantar mapa de na haonaid a scrúdaítear, a bheidh de dheasca sin suite i spás le buntomhas beag, de ghnáth a dó.

I dteannta leis an mapa seo féadtar struchtúr ordlathach i riocht chrainn a úsáid, struchtúr crannúil, a thugann dearcadh breise atá comhlántach agus comhghaolmhar.

*Ón 1.10.2004: Roinn na Ríomhaireachta, Royal Holloway University of London, Egham, Surrey TW20 0EX

Tús na hanailíse ná tacar de n oibiachta nó nithe a shonraítear, agus tacar de m tréitheanna. Cinntear ar luachanna (samhail-luachanna) ag crosbhealach gach oibiacht agus gach tréith.

Déanann pointe nó veictéoir san spás $\mathbb{R}^m$ ionadaíocht do gach oibiacht. Ar an dóigh chéanna tá pointe nó veictéoir san spás $\mathbb{R}^n$ ina ionadaí do gach tréith.

Maoinítear an dá spás sin (spásanna déacha a thugtar orthu) leis an méadracht χ^2. Go neamh-fhoirimiúil féadtar a rá gur méadracht Eoiclídeach le meáchain bhreise í an mhéadracht χ^2. Déantar mapáil den mhéadracht χ^2 seo i méadracht Eoiclídeach chun an mapa de na hoibiachta agus an ag am gcéanna de na tréitheanna a fháil, trí mheán factóir a shonrú. Tá tuilleadh eolais le fáil sna tagairtí seo a leanas: Benzécri (1976, 1992), Murtagh (2004).

2 Réamhchéimeanna na hAnailíse

Dúshraith na hoibre seo ná *méid tarlaithe* gach focal. Sainmhínímid *focal* mar shraith litreach, scartha le litir fhollamh nó le chomhartha poncaíochta ó sraith litreach at bith eile.

Ní réitíonn an beartas sin leis an dearcadh coitianta. Tosach na hoibre de ghnáth ná fréamhacha na bhfocal. Ins an ngnáth-dhearcadh go hiondúil glantar na focail atá an-mhinic ("is", "tá", 7rl) as an anailís.

Is cuid thabhachtach afách de feidhmiú ghaol-anailíse úsáid na bhfoirmfhocal seo. Tá siad go héifeachtach nuair atá éagsúlacht údar difriúla le haimsiú, nó difríochtaí i saothar aon údair amháin. I gcásanna áirithe is ionadaí iad foirmfhocail do focail theibí nach bhfuil teacht orthu chomh minic sin san téacs. Is deacair úsáid a bhaint as focail atá gann nó tearc, toisc nach bhfuil siad ionchompáirideach dá réir. Dearcadh "mála focal" a thugtar ar an chur chuige anailíse atá againn anseo, nuair atá tosaíocht tugtha mar seo do húsáid na bhfoirmfhocal.

Coimeádaimid chomh maith gach focal mar atá sé, gan mar shampla fréamhacha na bhfocal a aimsiú nó foirmeacha difriúla na mbriathar a chaighdeánú. I saothar Bhenzécri (Murtagh agus Gopalan, 2004; Murtagh, 2004) tá traidisiún saibhir ann ina bhfuil léirléamh torthúil déanta ar théacsanna agus ar dhoiciméid sa bhFraincis, sa Spáinnis, sa Rúisis, san Araibis, sa nGréigis chlaisiceach agus nua-aimsearach, sa Laidin, 7rl.

Seo iad roinnt de na téacsanna gur thugamar fúthú go dtí seo:

- Úrscéalta de chuid Jane Austin
- Fabhalscéalta na ndearthár Grimm
- Tuairiscí timpiste eitleán
- Tuairiscí brionglóidí
- Na Catagóirí de chuid Arastotail

3 Na Téacsanna a Úsáideadh

Bainimid úsáid as dhá théacs de chuid Shéamais Mhic Ghrianna:

1. "An imirce", caibideal 15 den leabhar *Rann na Feirste* (1942).
2. "Miorbhail Cholum Cille", gearrscéal as *An Teach Nár Tógadh* (1948).

Tá fáil ar na téacsanna seo ag an seoladh www.smo.uhi.ac.uk/∼oduibhin/leigh. Níol siad sa Chaighdeán Oifigiúil ach ní haon bhac é sin dúinn toisc gur staidéar inmheánach atá idir lámha againn.

San dá théacs thuas-luaite tá 3040 agus 3591 focal, faoi seach. Le 6631 focal san iomlán, fuaireamar 2746 foirmfhocal ar leith.

D'fhonn comparáid a dhéanamh le teanganna eile, rinneamar scrúdú ar dhá fhabhalscéal de chuid na ndearthár Grimm, sa mBéarla agus sa nGearmáinis: "Das tapfere Schneiderlein", "The Brave Little Tailor"; agus "Der goldene Vogel", "The Golden Bird". Sa nGearmáinis bhí 2976 agus 2803 focal san dá théacs, agus 2874 foirmfhocal ar leith. Sa mBéarla bhí 3335 agus 3078 focal san dá théacs, agus 2218 foirmfhocal ar leith.

An cinneadh a dheinimid dá bharr ná nach mór an difríocht atá ann maidir le húsáid na bhfoirmfhocal sna teanganna aonair seo.

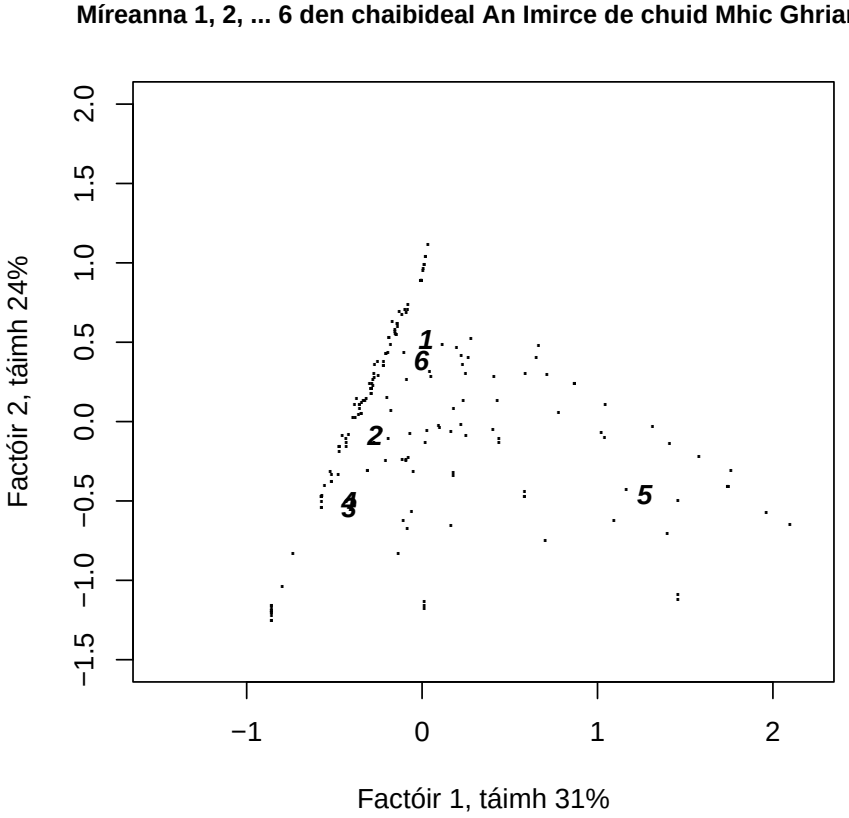
4 Anailís Dioscúrsa

Chun gluais nó druidim an téacs ó thaobh ábhair de a nochtadh, tabharfaimid faoi san tslí seo leanas.

Déanfaimid staidéar ar bonn gach roinn, nó chnuasacht altanna más gá, san téacs mar aonad bunúsach. Tugaimid *mír* ar gach aonad mar sin. I gcásanna áirithe ag baint le agallamh, mar shampla, nuair atá abairt amháin ina halt aonarach, cuirfimid abairtí le chéile chun alt a chruthú. Tá sé mar aidhm againn i gcás mar sin na míreanna a bheith a bheag nó a mhór cothram ó thaobh méid na bhfocal de.

Rinneadh gaol-anailís ar an tacar iomlán de 6 mír agus de 199 foirmfhocal ag a raibh minicíocht tarlaithe ≥ 3 san téacs ar fad. I spás na bhfactóir, tá gach aon mhír le fáil ag meánlár na bhfoirmfhocal i gcomhréir leis an bhaint eatarthu; agus tá sé den tsamhail chéanna do na foirmfhocail maidir leis na míreanna. Maidir leis an téacs "An imirce", tá an toradh le feiceáil i bhFigiúr 1.

I cló trom agus in iodáilaigh feictear na míreanna. Ar an ais chothrománach (factóir 1) feictear míreanna 1, 6, 2, 3, 4 ag druidim ón taobh deas i dtreo an taoibh chlé. Níl tábhacht ag baint le treoshuíomh na n-aiseanna. Tá mír 5 idir mhíreanna 3 agus 2, ach tá sí i bhfad ón druidim sin 1/6, 2, 4/3. Feicfear ar ball go bhfuil an téama anseo cuíosach neamh-ionann i gcomparáid leis an chuid eile. Ar an ais ingearach tá mír 5 scartha de na míreanna eile.



Figiúr 1: An príomhphlána atá mar thoradh ag an ghaol-anailís. Tá na míreanna léirithe i gcló trom; agus na foirmfhocail léirithe mar phointí.

Feicimid anois ar chinn de na foirmfhocail is suntasaí i bhFigiúr 1. Sa téacs "An imirce", tá sé mír ann, agus mar atá thuas-luaite bainimid úsáid as foirmfhocail le 3, nó níos mó ná 3, tharlú sa téacs ar fad.

Figiúr 1: druidim maidir leis na míreanna: 1 $\longrightarrow$ 2 $\longrightarrow$ 3/4 $\longrightarrow$ 5 ("contretemps") $\longrightarrow$ 6 atá a bheag nó a mhór cothram le 1.

Roghnú de foirmfhocail áisiúla agus baint acu leis na míreanna faoi leith:

1. Mír 1: Meiriceá, bás, paidir, phill.

   Míniú: *Deireadh an aistir – deireadh shaoil.*

2. Mír 2: bhaile, arais, aithne, chaoin.

   Míniú: *An áit a bhí tréigthe taobh thiar.*

3. Míreanna 3/4: feirste, rann, maidin, am, doras.

   Míniú: *Rann na Feirste agus Meiriceá an imircigh, fite fuaite ina chéile.*

4. Mír 5: teach, damhsa, iomlán, bainse, ceol.

   Míniú: *"Contretemps", ócáid na hoíche roimh ré.*

5. Mír 5: cumhaidh, athair, mháthair, bealach, cúl, sgáile, tocht. Míniú: *Comhbhrón and comhbhá leo siúd ar tí imeacht, agus leo siúd ag deireadh a saol.*

Sa téacs "Miorbhail Cholum Cille", tá cúig mhír ann, agus mar atá thuas-luaite bainimid úsáid as foirmfhocail le 3, nó níos mó ná 3, tharlú sa téacs ar fad.

Figiúr 2: druidim maidir leis na míreanna: 1 $\longrightarrow$ 2 $\longrightarrow$ 3 agus ansin saghas ath-thosnú le 4 $\longrightarrow$ 5. Ag an deireadh, 6 maraon le 1 ó thaobh teilgin ar ais 1 de.

Roghnú de na foirmfhocail áisiúla agus baint acu leis na míreanna faoi leith:

1. Mír 1: Cille, Cholum, Colum, suim, beatha, stócach.

   Míniú: *Cúlra faoin leabhar faoi Cholm Cille.*

2. Mír 2: bheatha, eallach, cruaidh, gabháil, ngartán, suaimhneas.

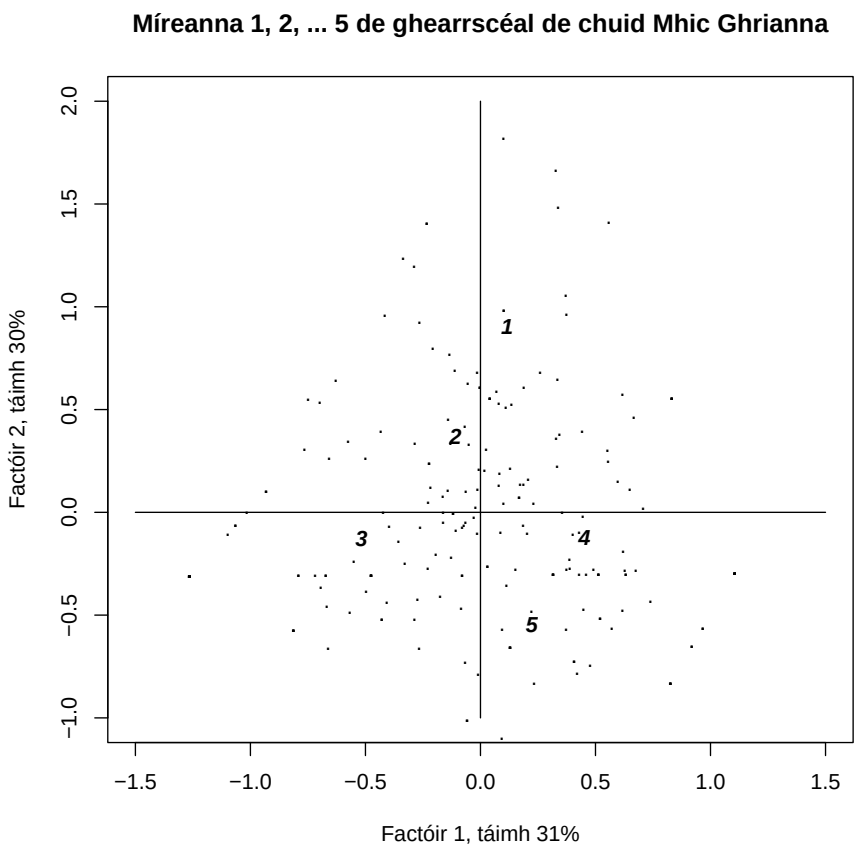
   Míniú: *Tús an chur síos ar an uncal agus ar saol phríomhphearsanta an scéil (Tarlach).*

3. Mír 3: leabhar, léigheamh, maidin, bocht, bpunta, cearrbhach, bhocht, chearrbhach, tráthnóna, aonach, lae, oibre, oidhche, shaoghal, bhó, dhíol, díol, leabaidh, luach, obair.

   Míniú: *Leathnú agus doimhniú ar chuntas saol Tharlach agus an uncail.*

4. Mír 4: phighinn, píobaire, airgead, daoine, fial, phócaí, scéal, scilling, seinm.

   Míniú: *Cathú Tharlach.*



Figiúr 2: Gearr-scéal "Miorbhail Cholum Cille". An príomhphlána atá mar thoradh ag an ghaol-anailís. Tá na míreanna léirithe i gcló trom; agus na foirmfhocail léirithe mar phointí.

5. Mír 5: uncal, phíobaire, uair, ádhbhar, spéir, dhíoghbháil, éadaigh.

   Míniú: *Neimisis, nach bhfuil ro-soiléar ó na focail anseo; ach tá leid tugtha dúinn i bhFigiúr 2 leis an "filleadh thar nais" maidir le ghluaiseacht na míreanna:* $1 \longrightarrow 2 \longrightarrow 3$ *ach ansin athrú tobann maidir le* $4 \longrightarrow 5$.

5 Cinneadh agus Focal Scoir

Ó na Figiúirí tá saghas chasadh thar nais, nó filleadh, ar ghnéithe a bhí dá bplé cheana féin le tabhairt faoi deara. Cuir i gcás 6 agus 1 i bhFigiúr 1; agus i bhFigiúr 2, tá 5 an-ghar do mhír 1 ó thaobh chomhordanáidí ar ais 1 de. Bhéafá ag súil le chasadh thar nais den tsaghas sin san fhilíocht nó san amhránaíocht. An freagra a chuirimid ná: an bhfuil casadh thar nais mar sin neamh-choitianta nó coitianta san litríocht agus san réasúnú go ginearálta?

An féidir linn mar sin úsáid a bhaint as teagmhas mar sin chun sórt chomhaontú barúla a fháil ar smaointe – ar loighic – an údair?

Déanfaimid freagraí mar sin a scrúdú amach anseo, ar bonn chur chuige atá idir chamáin ag Day agus McMorris (2003).

Tagairtí

1. J.P. Benzécri, *Analyse des Correspondances*, 2e ed., Dunod, 1976.

2. J.P. Benzécri, *Correspondence Analysis Handbook*, Marcel Dekker, 1992.

3. W.H.E. Day agus F.R. McMorris, *Axiomatic Consensus Theory in Group Choice and Biomathematics*, Society for Industrial and Applied Mathematics, 2003.

4. F. Murtagh agus T.K. Gopalan, "Content analysis of text: the correspondence analysis approach", faoi léirmheas ag *Artificial Intelligence Review*, 2004.

5. F. Murtagh, *Correspondence Analysis and Data Coding with Java and R*, dá scríobh, Chapman and Hall/CRC Press, 2004.