Dealing with ubiquitous data

Michael Cysouw

Linguists amass data …

- Typological databases (WALS)
- Lexical databases (IDS, LWT)
- Textual material (DoBeS, Rausing)

… and there is even more available

- Full-text searching of grammars/dictionaries
- (Online) electronic dictionaries
- The web itself as a corpus

Although there are still some problems ...

- Management of sources
- Orthography
- Terminology

How can we use all this data for language comparison ?

Example 1

Typology from language-particular similarities

Inchoative - causative verb pairs

- Inchoative
  "The stick broke."

- Causative
  "The girl broke the stick."

Based on data from: Haspelmath, Martin. 1993. "More on the typology of inchoative/causative verb alternations." In: Comrie, B. & Polinsky, M. (eds.) *Causatives and transitivity*. Amsterdam: Benjamins, 87-120.

Verb pairs investigated

begin	dry	melt
boil	fill	open
break	finish	rise/raise
burn	freeze	rock
change	gather	roll
close	get lost/	sink
connect	lose	split
destroy	go out/put	spread
develop	out	stop
die/kill	improve	turn
dissolve	learn/teach	wake up

Arabic

Class A: C / CC
1. saʿaa / saʿʿaa
8. darasa / darrasa
14. damara / dammara
31. waqafa / waqqafa

Class B: in / ø
2. inkasara / kasara
5. infataʿa / fataʿa
6. inqafala / qafala
13. inṣahara / ṣahara
30. inšaqqa / šaqqa

Class C: t / ʔ
3. iʿtaraqa / ʔaʿraqa
22. intahaa / ʔanhaa

Class D: t / ø
9. iltamma / lamma
10. intašara / našara
17. irtabaṭa / rabaṭa
21. irtafaʕa / rafaʕa
27. imtalaʔa / malaʔa

Class E: ø / ʔ
11. ġariqa / ʔaġraqa
18. ġalaa / ʔaġlaa
23. daara / ʔadaara
26. ðaaba / ʔaðaaba

Class F: ta / ø
12. tabaddala / baddala
16. taṭawwara / ṭawwara
19. taʔarjaʿa / ʔarjaʿa
24. tadaʿraja / daʿraja
25. tajammada / jammada
28. taʿassana / ʿassana

Singular cases:
4. maata / qatala
7. badaʔa
15. daaʕa / xasira
20. inṭafaʔa / ʔaṭfaʔa
29. jaffa / jaffafa

English

Class A: Identical
1. wake up
2. break
3. burn
5. open
6. close
7. begin
9. gather
10. spread
11. sink
12. change
13. melt
16. develop
17. connect
18. boil
19. rock
22. finish
23. turn
24. roll
25. freeze
26. dissolve
27. fill
28. improve
29. dry
30. split
31. stop

Singular cases:
4. die / kill
8. learn / teach
14. be destroyed / destroy
15. get lost / lose
20. go out / put out
21. rise / raise

Finnish

Class A: ø / ttA
1. herätä / herättää
3. palaa / polttaa
8. oppia / opettaa
10. levitä / levittää
13. sulaa / sulattaa
18. kiehua / kiehuttaa
19. kiikkua / kiikuttaa
20. sammua / sammuttaa
21. kohota / kohottaa
22. loppua / lopettaa
24. vieriä / vierittää
25. jäätyä / jäädyttää
26. liueta / liuottaa
31. pysähtyä / pysähdyttää

Class B: U / A
2. murtua / murtaa
12. muuttua / muuttaa
16. kehittyä / kehittää
23. vääntyä / vääntää
27. täyttyä / täyttää
28. parantua / parantaa

Class C: UtU / ø
5. avautua / avata
6. sulkeutua / sulkea
14. tuhoutua / tuhota

Singular cases:
4. kuolla / tappaa
7. alkaa / aloitaa
9. kokoontua / koota
11. laskea
15. hukkaantua / hukata
17. yhtyä / yhdistää
29. kuivaa / kuivata
30. haljeta / halkaista

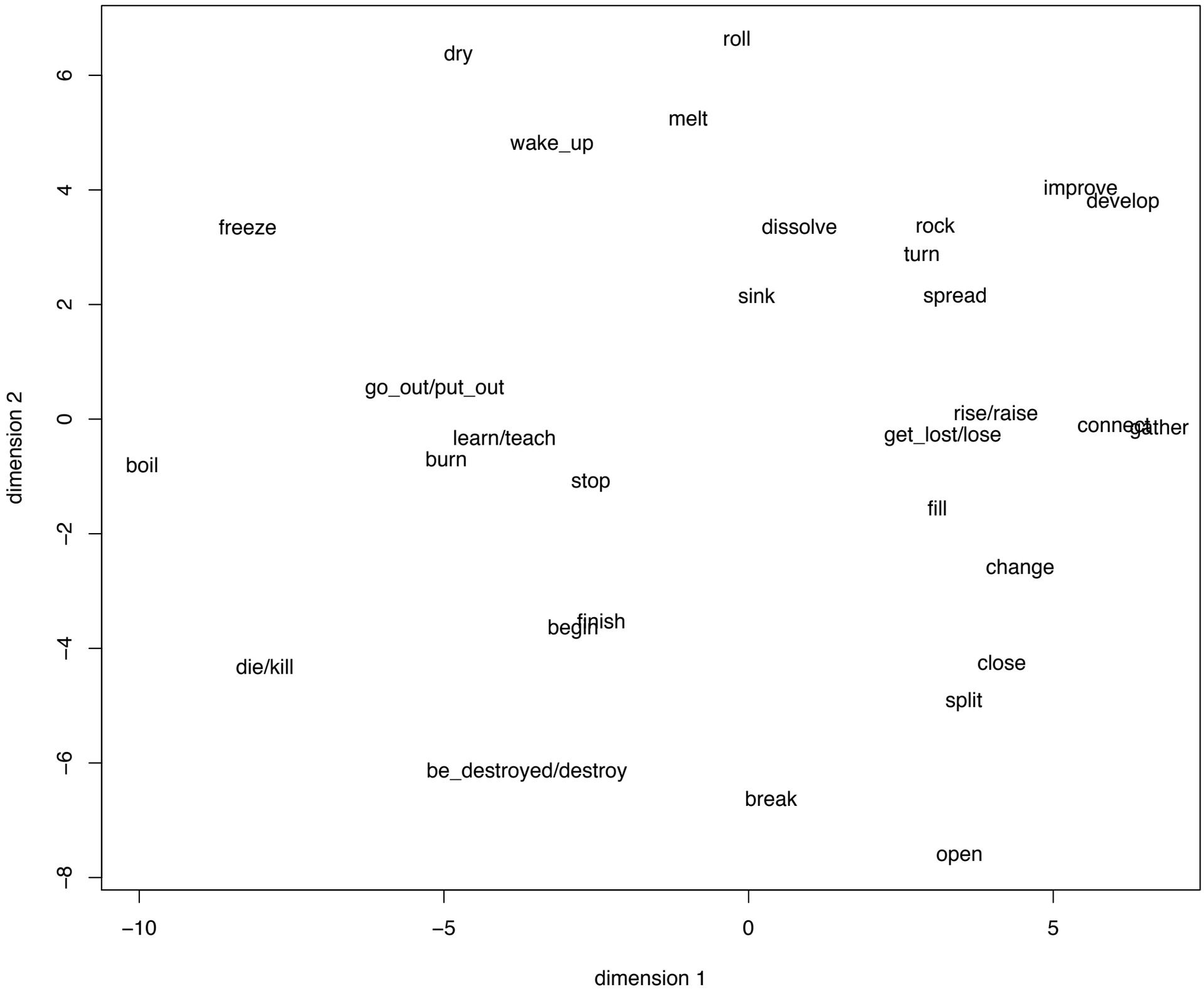
French

Class A: se / ø
1. se réveiller / réveiller
2. se briser / briser
5. s’ouvrir / ouvrir
6. se fermer / fermer
9. s’assembler / assembler
10. s’étendre / étendre
11. s’enfoncer / enfoncer
15. se perdre / perdre
16. se développer / développer
17. se lier / lier
19. se balancer / balancer
20. s’éteindre / éteindre
21. se lever / lever
23. se tourner / tourner
26. se dissoudre / dissoudre
27. se remplir / remplir
28. s’améliorer / améliorer
30. se fendre / fendre
31. s’arrêter / arrêter

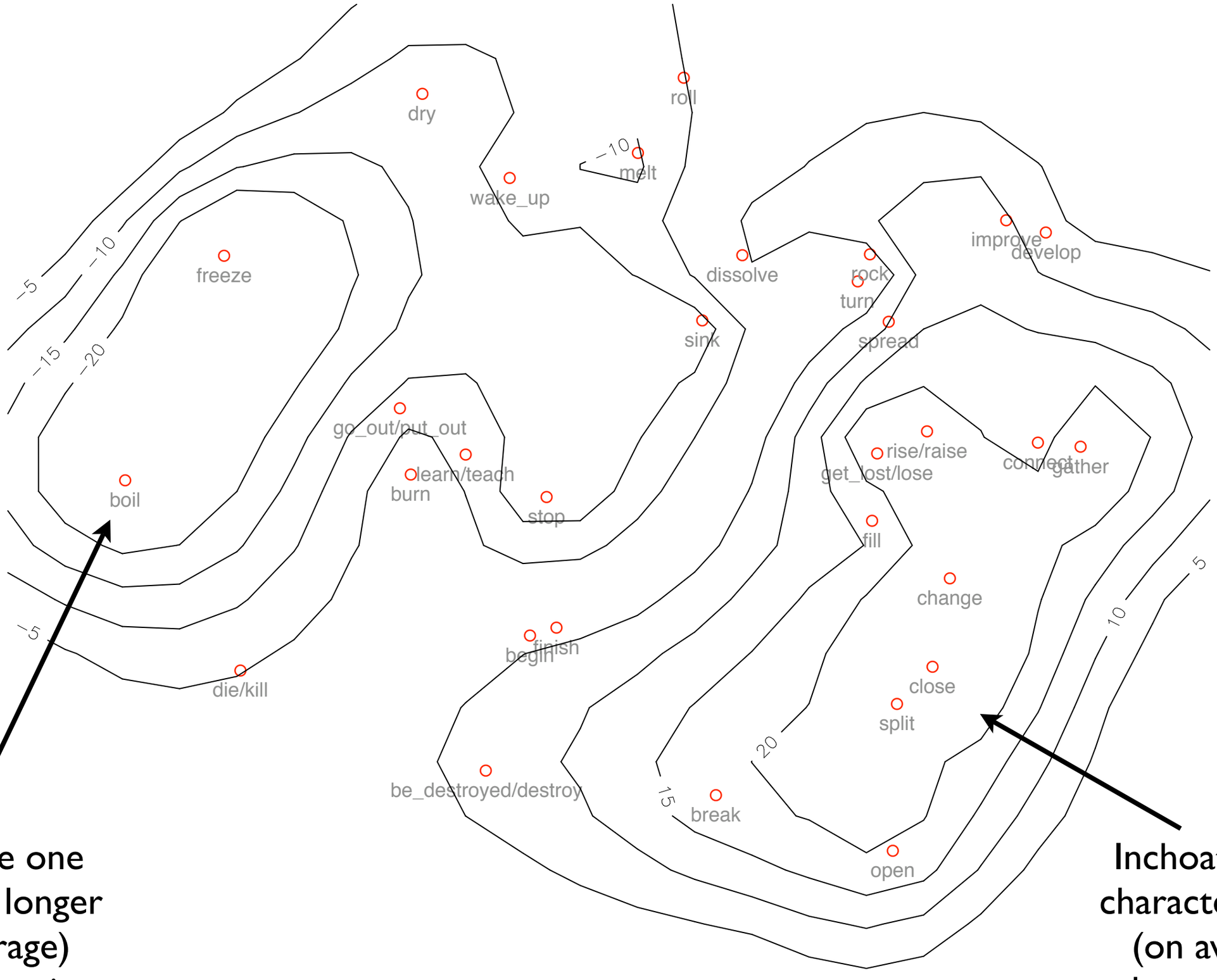
Class B: Identical
3. brûler
7. commencer
8. apprendre
12. changer
22. finir
24. rouler
25. geler
29. sécher

Class C: ø / faire
13. fondre / faire fondre
18. bouillir / faire bouillir

Singular cases:
4. mourir / tuer
14. être détruit / détruir


roll
dry
melt
wake_up
improve
develop
freeze
dissolve
rock
turn
sink
spread
go_out/put_out
rise/raise
connect
gather
learn/teach
get_lost/lose
boil
burn
stop
fill
change
finish
begin
die/kill
close
split
be_destroyed/destroy
break
open
6
4
2
0
−2
−4
−6
−8
−10
−5
0
5
dimension 2
dimension 1

Inchoative–Causative Character Count



Causative one character longer (on average) than inchoative

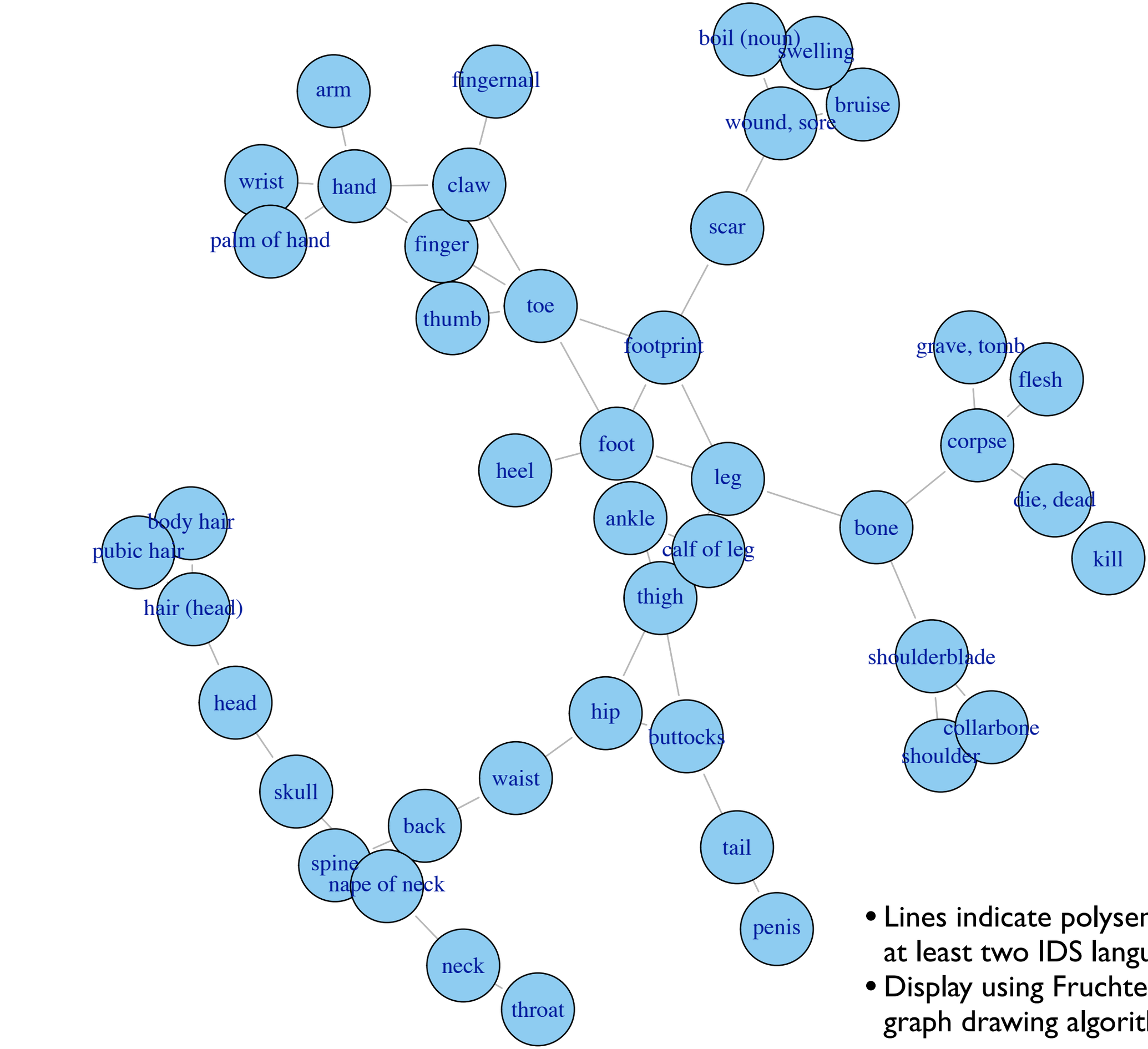
Inchoative one character longer (on average) than causative

Example 2

Semantics from recurrent polysemy

using the Intercontinental Dictionary Series

chapter_id	entry_id	trans_english	Aché [Phonemic]	Aguaruna [Phonemic]	Ayoreo [Phonemic]
4	110	body	eče	iyaš	gaay; 'uud-ode?
4	120	skin, hide	pire	duwap-ɨ pušihi	a'ooi
4	130	flesh	o?o	nɨhɨ	a'ooi; a'ragaⁿua
4	140	hair (head)	a?a; tõ a?a	intaš	gaa'terero
4	142	beard	buta	intaš	abu'haa-sii
4	144	body hair	eče a?a	susu	o'hoi
4	145	pubic hair	ta a?a		
4	146	dandruff		du'sɨ buuk	
4	150	blood	bu-ɨ?ɨ	numpa	i'yoi
4	151	vein, artery	ga-ɨǰɨ	num'pa hinti	'ku-kura-y
4	160	bone	ikã	ukunč	aⁿokeei
4	162	rib	rukã	pagaɨ	orotabi'di
4	170	horn	ačĩ	kaču	u'ei
4	180	tail	bua?a	uhu'kɨ	ka'ri
4	190	back	pɨče	tuntup; maya-tai	giido'booi
4	191	spine	pɨče ikã	tagkihi	oo'hi
4	200	head	tõ?õ	buuk	gaa'toi
4	201	side of head, temple	tõ?õ ẽbe	kahašik	
4	202	skull	tõ gape	buuk sakahu	
4	203	brain	tõ?õ	buɛuk	ta'rooi



- Lines indicate polysemy in at least two IDS languages
- Display using Fruchterman-Reingold graph drawing algorithm

Example 3

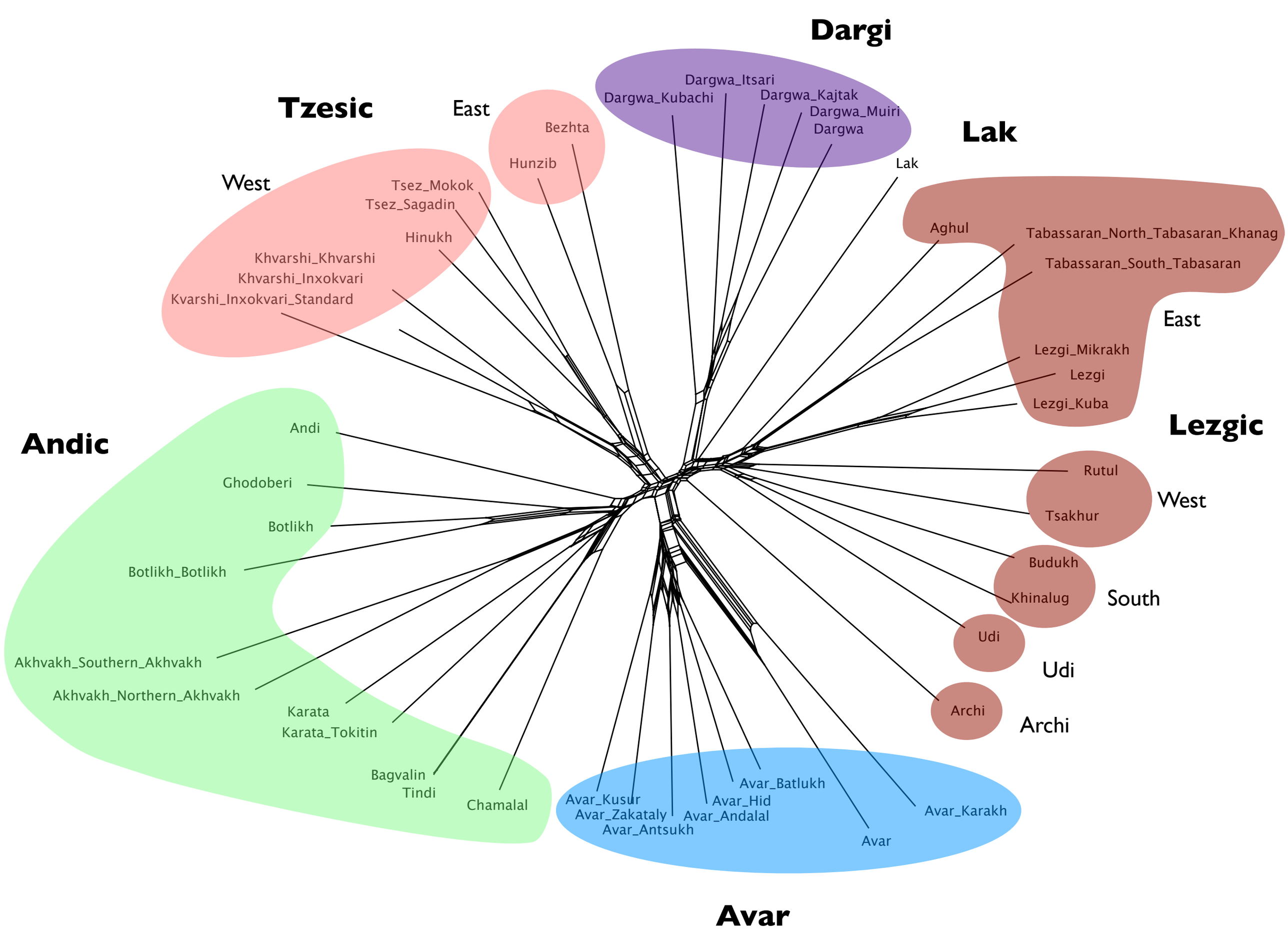
Language similarities by counting word-edits

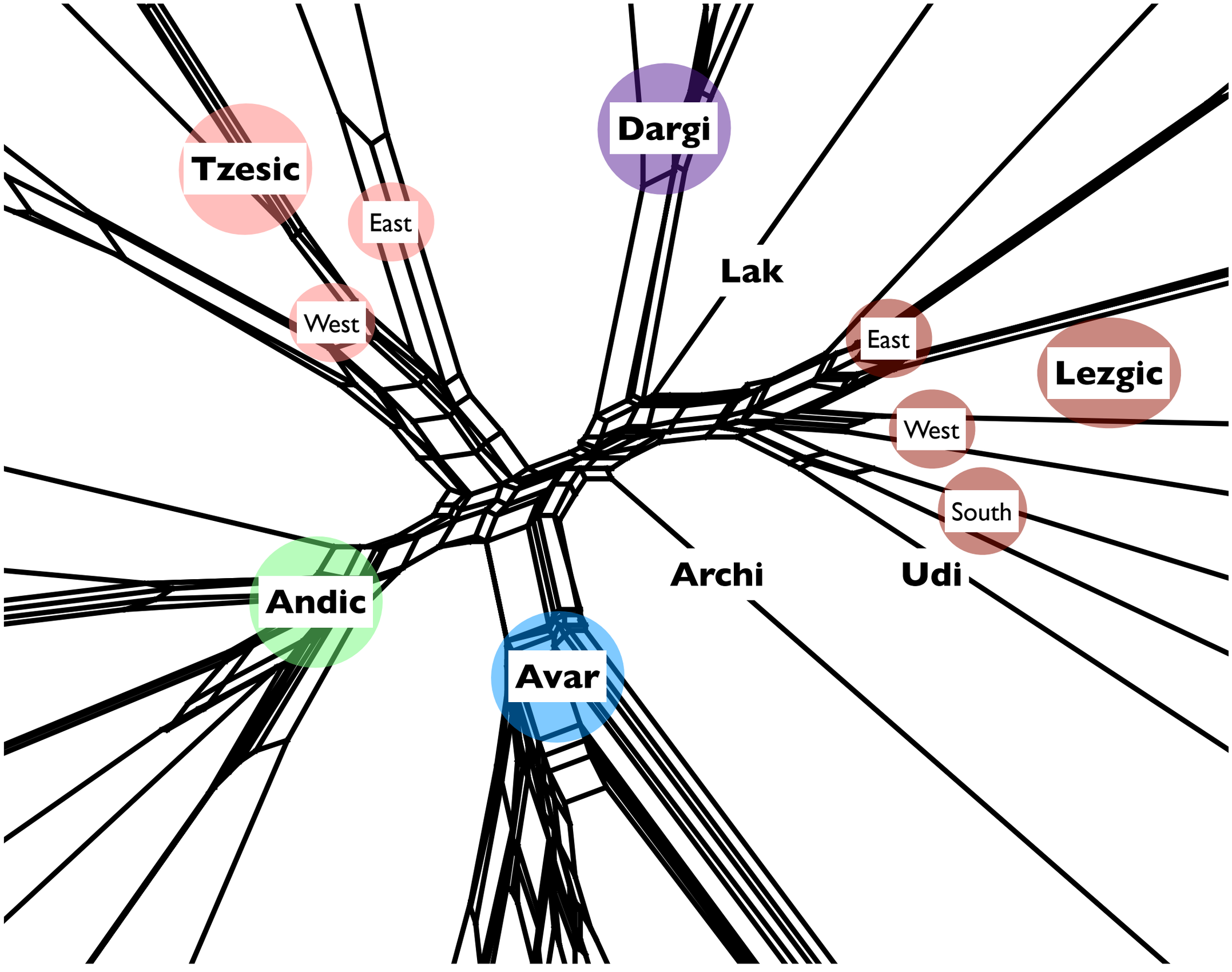
using the Intercontinental Dictionary Series

abcdefgh ⟷ bacdxfgh

one addition	abcdefgh	abcd**x**efgh
one deletion	abcdx**e**fgh	abcdxfgh
one inversion	abcdxfgh	**ba**cdxfgh

Total of three changes needed

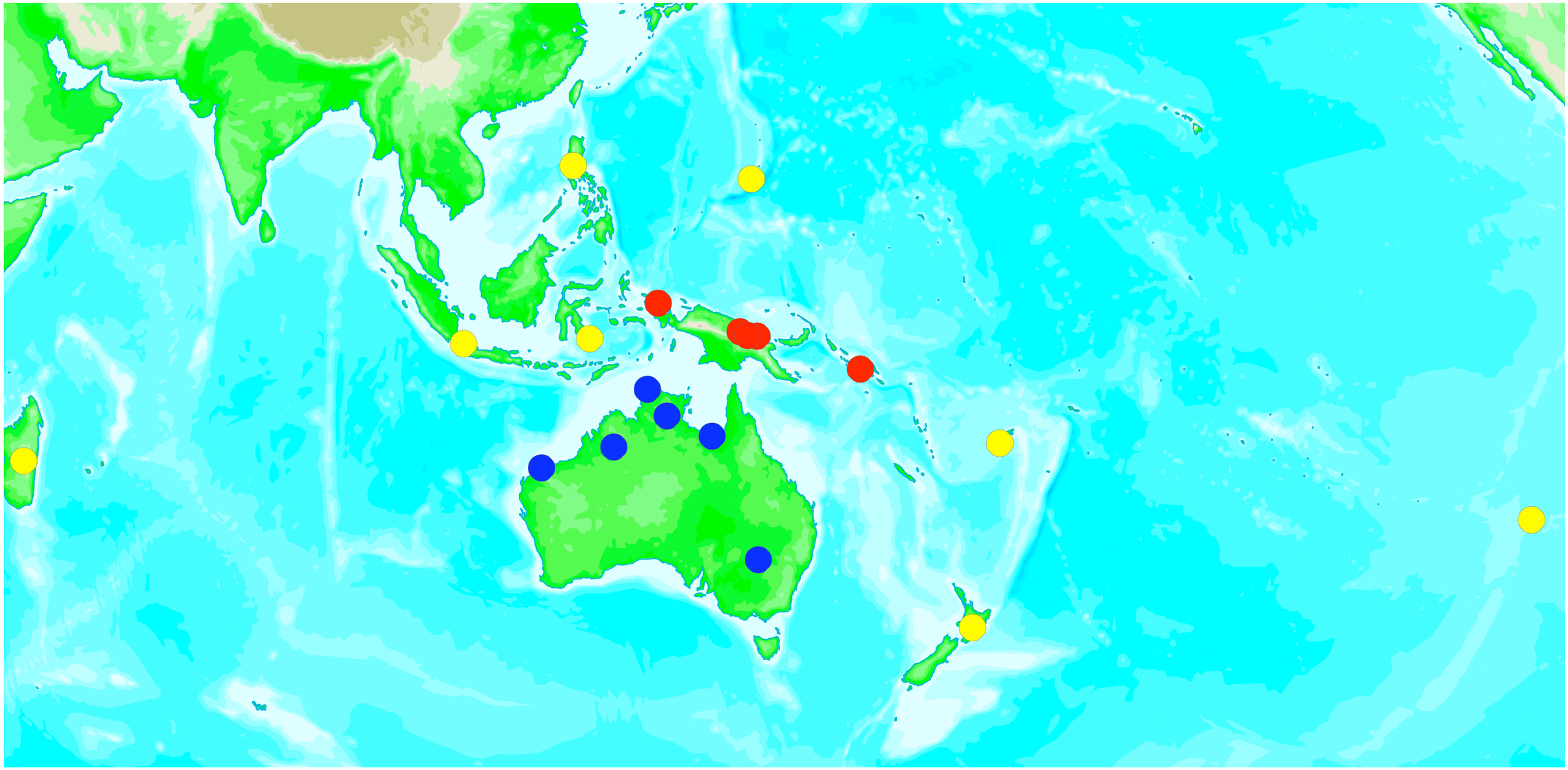

Dargi
Tzesic
East
Dargwa_Itsari
Dargwa_Kubachi
Dargwa_Kajtak
Dargwa_Muiri
Dargwa
Bezhta
Hunzib
Lak
Lak
West
Tsez_Mokok
Tsez_Sagadin
Hinukh
Khvarshi_Khvarshi
Khvarshi_Inxokvari
Kvarshi_Inxokvari_Standard
Aghul
Tabassaran_North_Tabasaran_Khanag
Tabassaran_South_Tabasaran
East
Lezgi_Mikrakh
Lezgi
Lezgi_Kuba
Lezgic
Andic
Andi
Rutul
Tsakhur
West
Budukh
Khinalug
South
Udi
Udi
Archi
Archi
Ghodoberi
Botlikh
Botlikh_Botlikh
Akhvakh_Southern_Akhvakh
Akhvakh_Northern_Akhvakh
Karata
Karata_Tokitin
Bagvalin
Tindi
Chamalal
Avar_Batlukh
Avar_Kusur
Avar_Hid
Avar_Zakataly
Avar_Andalal
Avar_Antsukh
Avar_Karakh
Avar
Avar

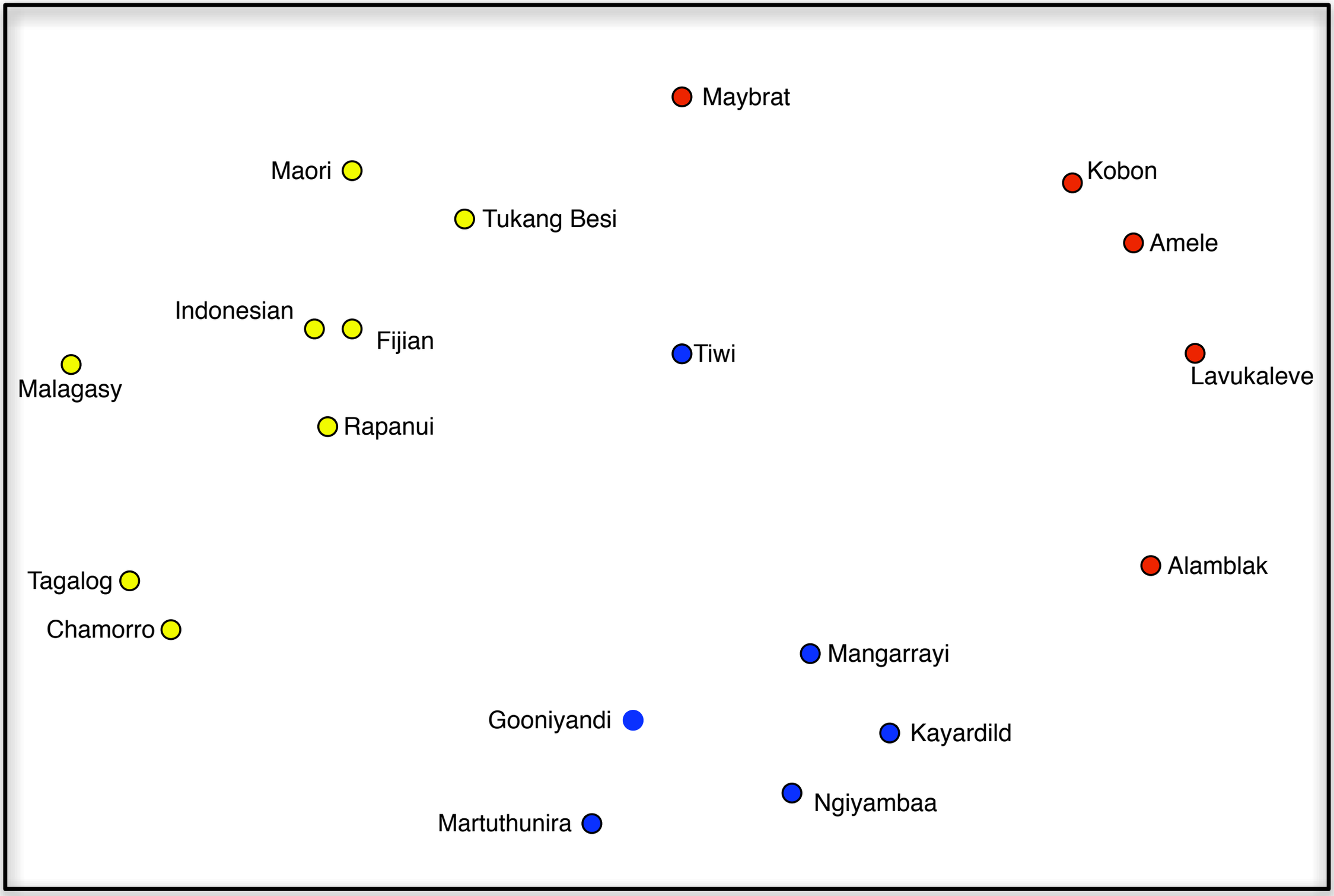

Tzesic
East
West
Dargi
Lak
East
Lezgic
West
South
Udi
Archi
Andic
Avar

Example 4

Language similarities

using the World Atlas of Language Structures




Maybrat
Maori
Kobon
Tukang Besi
Amele
Indonesian
Fijian
Tiwi
Lavukaleve
Malagasy
Rapanui
Alamblak
Tagalog
Chamorro
Mangarrayi
Gooniyandi
Kayardild
Ngiyambaa
Martuthunira

Maybrat

Kobon

Tukang Besi

Amele

Tiwi

Lavukaleve

Alamblak

Mangarrayi

Gooniyandi

Kayardild

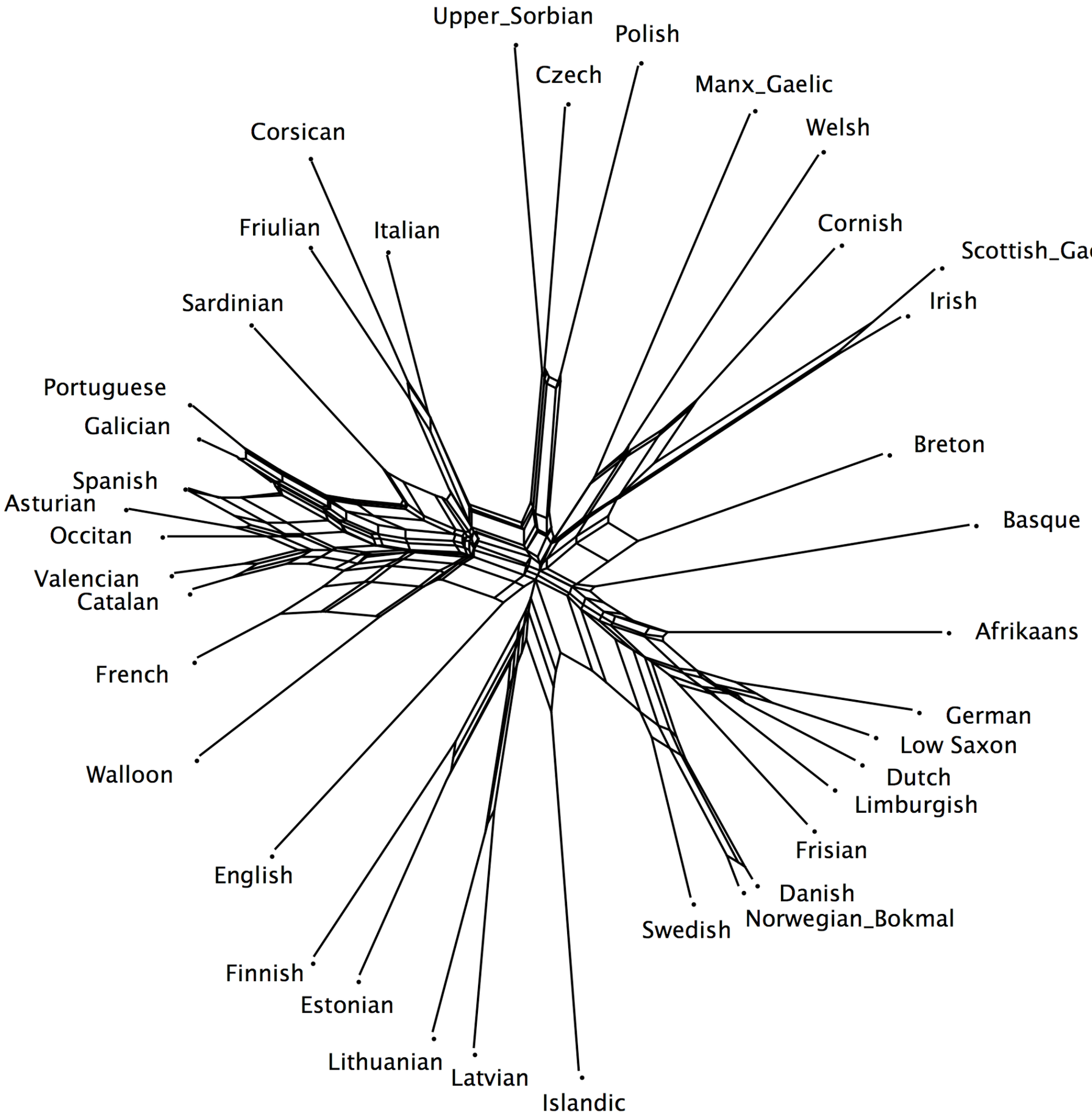
Ngiyambaa

Martuthunira

Example 5

Language similarities by graphotactics

using unannotated web-corpora


Upper_Sorbian
Polish
Czech
Manx_Gaelic
Welsh
Corsican
Cornish
Friulian
Italian
Scottish_Gaelic
Irish
Sardinian
Portuguese
Galician
Breton
Spanish
Asturian
Basque
Occitan
Valencian
Catalan
Afrikaans
French
German
Low Saxon
Walloon
Dutch
Limburgish
Frisian
English
Danish
Norwegian_Bokmal
Swedish
Finnish
Estonian
Lithuanian
Latvian
Islandic

The future ?!

- Preparation of data is still very laborious
- Applicability of quantitative methods and visualisations has to be explored for linguistics
- Suitable approaches have to be adapted for linguistic data