Areality in colexification patterns

Maria Koptjevskaja-Tamm,[1] Volker Gast[2]

[1]Stockholm, [2]Jena & Olomouc

28 June 2018

Introduction

- General context: Areality in lexical typology (e.g. Koptjevskaja-Tamm and Liljegren 2017)[1]
- Large-scale convergence patterns in the organization of the lexicon, with a focus on *colexification*.[2]
- Using data from two major lexical databases,
  - The Crosslinguistic Database of Colexifications (CLICS),[3] and
  - the Automated Similarity Judgment Programme (ASJP).[4]

[1] M. Koptjevskaja-Tamm and H. Liljegren (2017). "Semantic patterns from an areal perspective". In: *The Cambridge Handbook of Areal Linguistics*. Ed. by R. Hickey. Cambridge: Cambridge University Press, pp. 204–236.

[2] A. François (2008). "Semantic maps and the typology of colexification: Intertwining polysemous networks across languages". In: *From Polysemy to Semantic change: Towards a Typology of Lexical Semantic Associations*. Ed. by M. Vanhoeve. Amsterdam: Benjamins.

[3] J.-M. List et al. (2014). *CLICS: Database of Cross-Linguistic Colexifications*. Version 1.0. http://CLICS.lingpy.org. Marburg: Forschungszentrum Deutscher Sprachatlas.

[4] S. Wichmann, E. Holman, and C. H. Brown (2016). *The ASJP Database*. Version 17.

Overview

Areal lexico-semantic patterns

- Types of areal lexico-semantic patterns:
  - lexico-semantic parallels (colexification patterns, lexico-constructional patterns),
  - shared formulaic expressions,
  - area-specific lexicalizations.
    (cf. Koptjevskaja-Tamm and Liljegren 2017):[5]
- Colexification pairs as the 'atomic' concept of (lexical) semantic maps.

[5] M. Koptjevskaja-Tamm and H. Liljegren (2017). "Semantic patterns from an areal perspective". In: *The Cambridge Handbook of Areal Linguistics*. Ed. by R. Hickey. Cambridge: Cambridge University Press, pp. 204–236.

The data

- **CLICS**: 92,805 data points
- **ASJP**: 2,060,856 data points
- Each data point is a quadruple of the form $<C_1, C_2, L, B>$ (C: concept, L: language, B: Boolean, TRUE or FALSE)
- Example: <'arm', 'hand', 'russ1263', TRUE>
- Some negative evidence could be inferred.
- Note that the concept of a binary distinction between colexification and differentiation is of course simplifying (e.g. 'arm', 'hand' in Russian).

The starting point

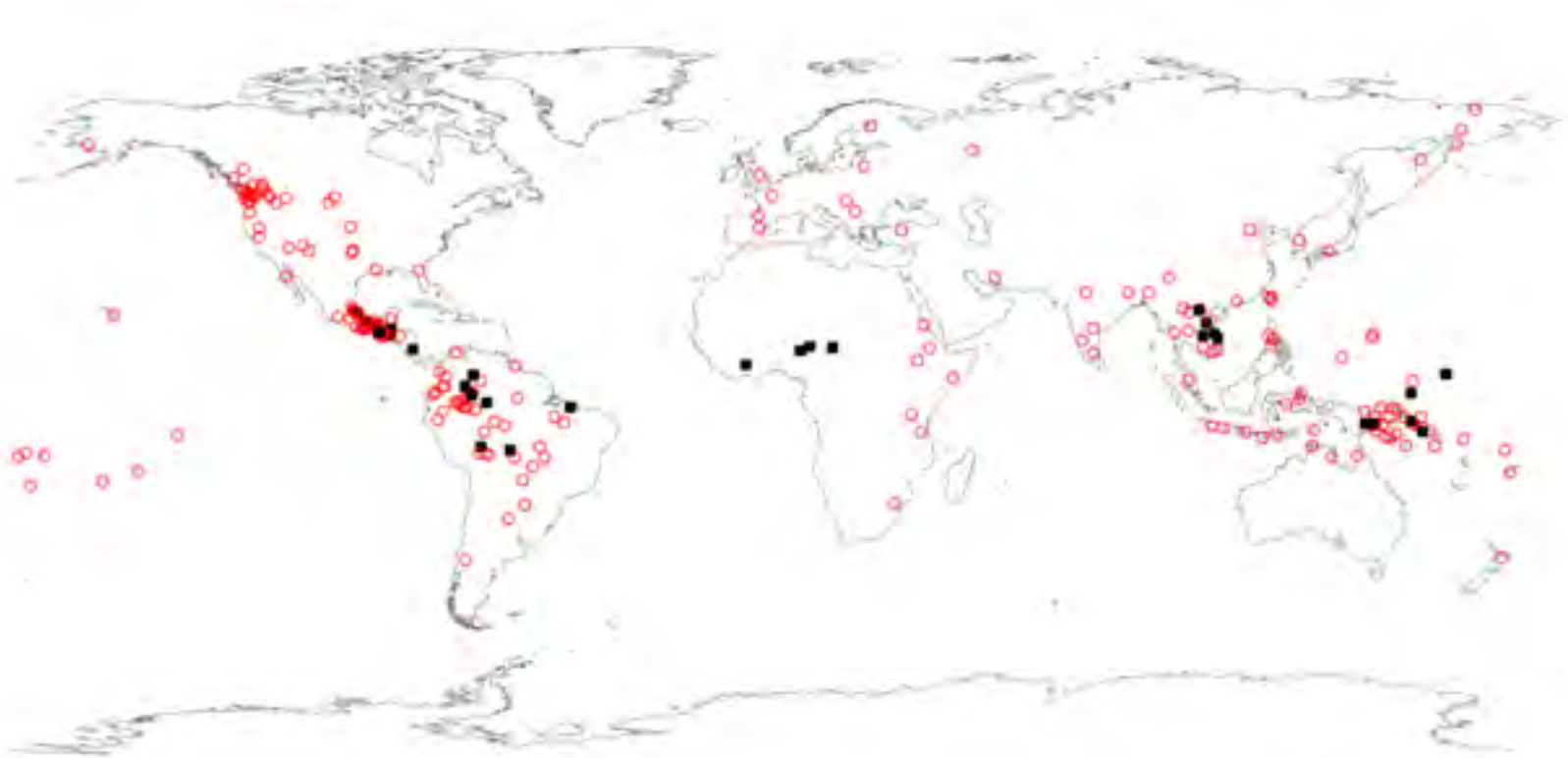


Figure 1: Colexification (black squares) vs. differentiation (red circles) of 'feather' and 'hair' in the ASJP-data

Identifying geographical clusters

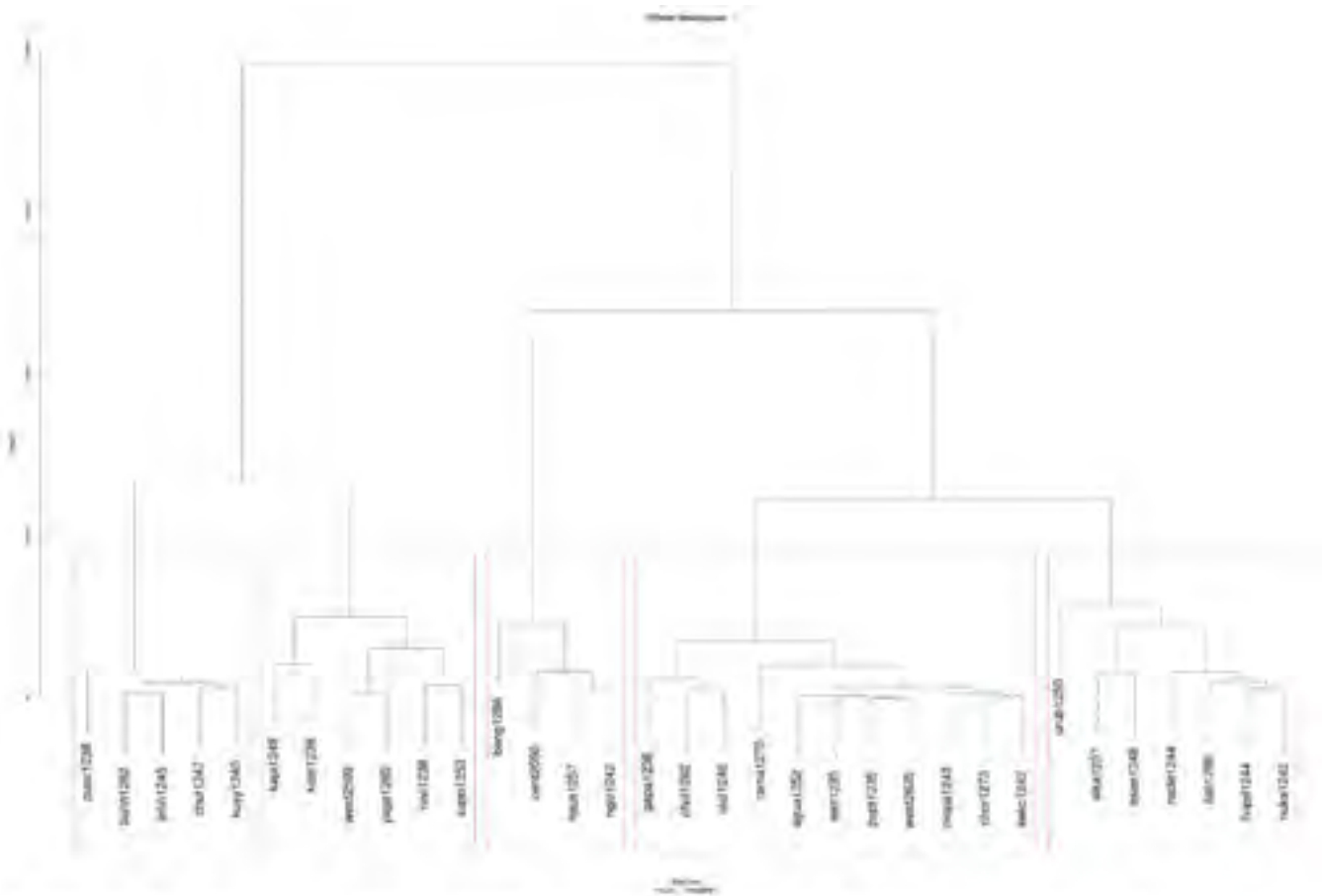


Figure 3: Dendrogram for the colexification pair <'feather', 'hair'> (meso-clusters)

Identifying geographical clusters

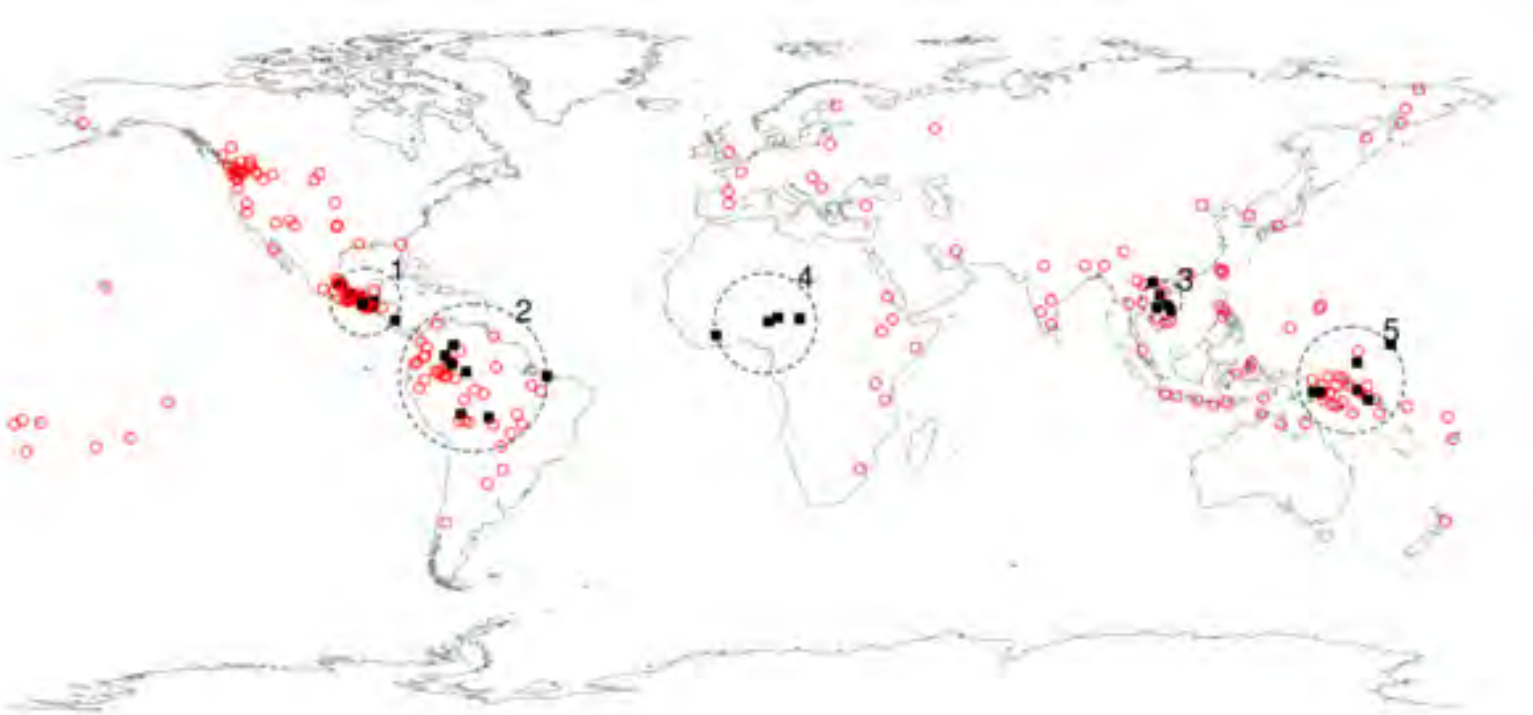


Figure 4: Hypothesized meso-clusters for the colexification pattern <'feather', 'hair'>

Testing the clusters

- For each cluster, a (Bayesian) mixed effects model was fitted (cf. Hadfield 2010,[6] Levshina 2010).[7]
- Language family was treated as a random effect.
- Moreover, a genealogical diversity index was calculated for each cluster (entropy).

[6] J. Hadfield (2010). "MCMC methods for Multi-response Generalised Linear Mixed Models: The MCMCglmm R Package". In: *Journal of Statistical Software* 33.2, pp. 1–22.

[7] N. Levshina (2015). *Bayesian logistic models with MCMCglmm: A brief tutorial*. http://www.natalialevshina.com/Documents/MCMCglmm_Tutorial.pdf.

Two clusters

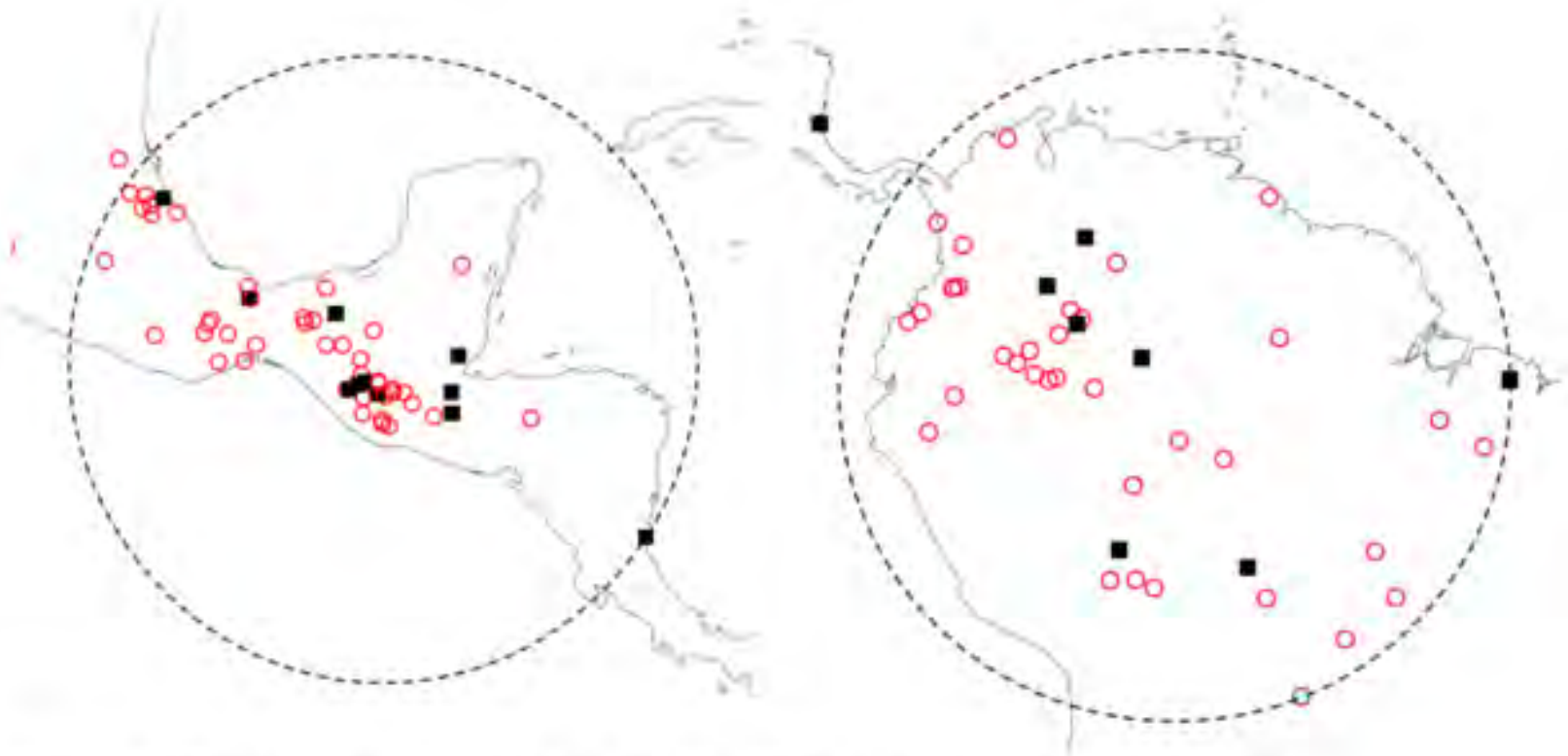


Figure 5: Cluster Areas 1 and 2 for the colexification pair <'feather', 'hair'>

Genealogical diversity in the clusters

Figure 6: Clusters 1 and 5 for <'feather', 'hair'>

Some results: ASJP

- We identified the clusters that were 'strongest' in terms of the regression statistics and genealogical diversity.
- Strongest areal clusters for the ASJP-data:
  - <fire, tree>
  - <mountain, stone>
  - <ear, leaf>
  - <bark, skin>

<fire, tree>

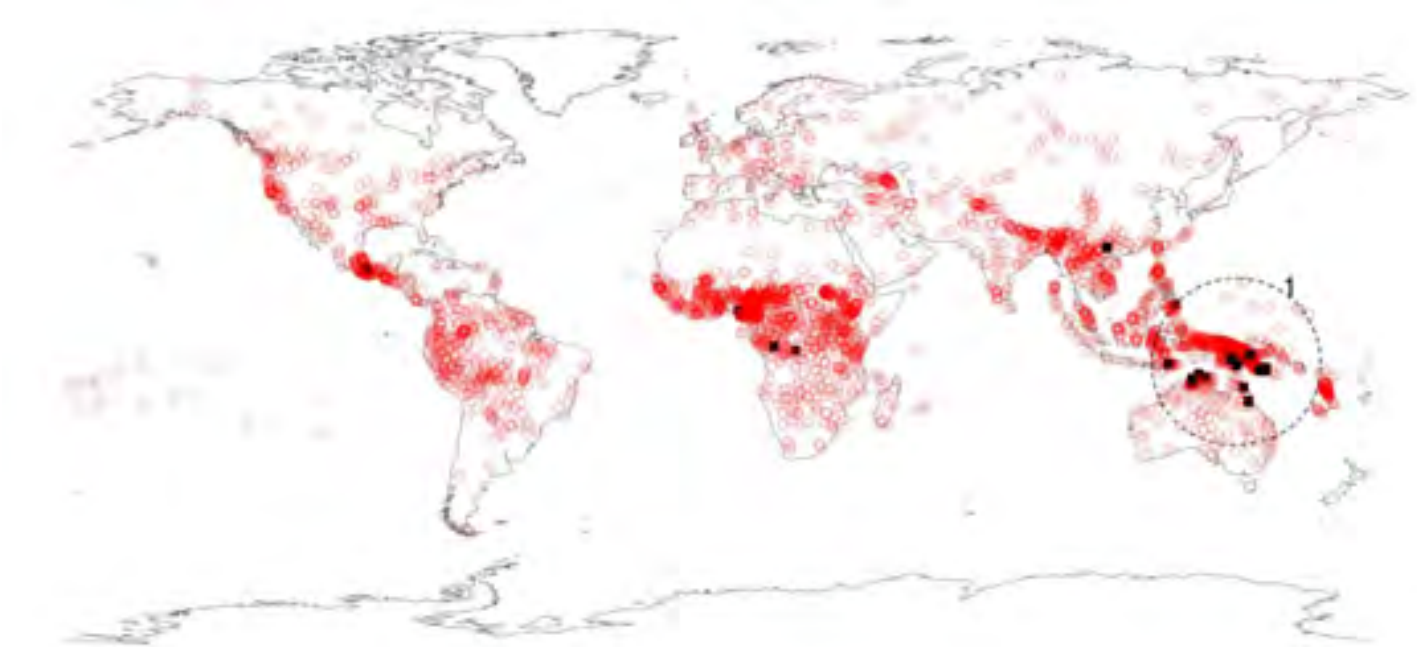


Figure 8: Significant meso-cluster for <'fire', 'tree'>

<fire, tree>: Discussion

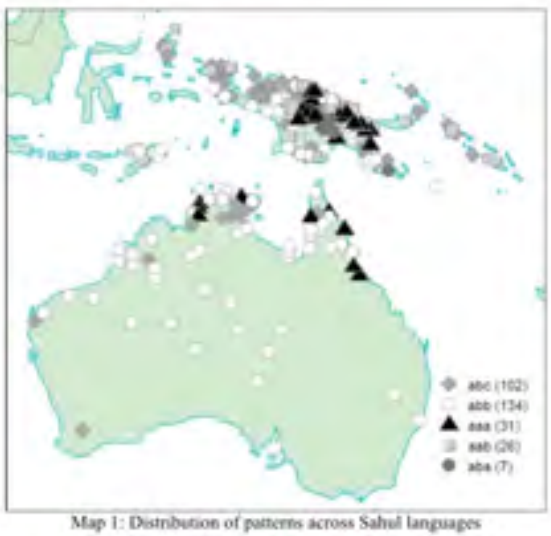
- Lexical-semantic relations between 'fire' and 'tree' is a well-known feature of the languages of Australia, New Guinea and the surrounding islands, cf. Schapper et al. (2016),[8] Östling (2016).[9]

[8] A. Schapper, L. S. Roque, and R. Hendery (2016). "Tree, firewood and fire in the languages of Sahul". In: *The Lexical Typology of Semantic Shifts*. Ed. by P. Juvonen and M. Koptjevskaja-Tamm. Berlin: de Gruyter Mouton, pp. 355–422.

[9] R. Östling (2016). "Studying colexification through massively parallel corpora". In: *The Lexical Typology of Semantic Shifts*. Ed. by P. Juvonen and M. Koptjevskaja-Tamm. Berlin: de Gruyter Mouton, pp. 157–176.

<fire, tree> in Sahul

Schapper et al. (2016)



Map 1: Distribution of patterns across Sahul languages

ASJP

Östling (2016)

Schapper et al. (2016): Frequencies

- Most, but not all, languages colexifying 'tree' and 'fire' also exibit colexification of both concepts with 'firewood' (31/38) (which is not on the Swadesh list).

Table 6: Occurrence of patterns in Sahul

		Papuan languages	Australian languages	TOTAL	Percent
abc	full differentiation	92	10	102	34%
abb	firewood/fire colexification	70	64	134	45.5%
aab	tree/firewood colexification	26	0[8]	26	8.5%
aba	firewood differentiation	7	0	7	2%
aaa	full colexification	22	9	31	10%

Schapper et al. (2016): Frequencies

- 'Subcolexification': the 'primary lexifier' is the same, e.g. Daga *oma* 'tree, fire', *oma oaewa* 'firewood'.
- The number of patterns is reduced.

Table 18: Reclassification of languages on the basis of subcolexifications

		Papuan languages	Australian languages	Patterns of subcolexification
abc	full differentiation	66	2	--
abb	firewood/fire colexification	80	52	aBb, abB
aab	tree/firewood colexification	33	2	aAb
aaa	full colexification	38	7	aAa, aAA, aaA

<fire, tree>: Discussion

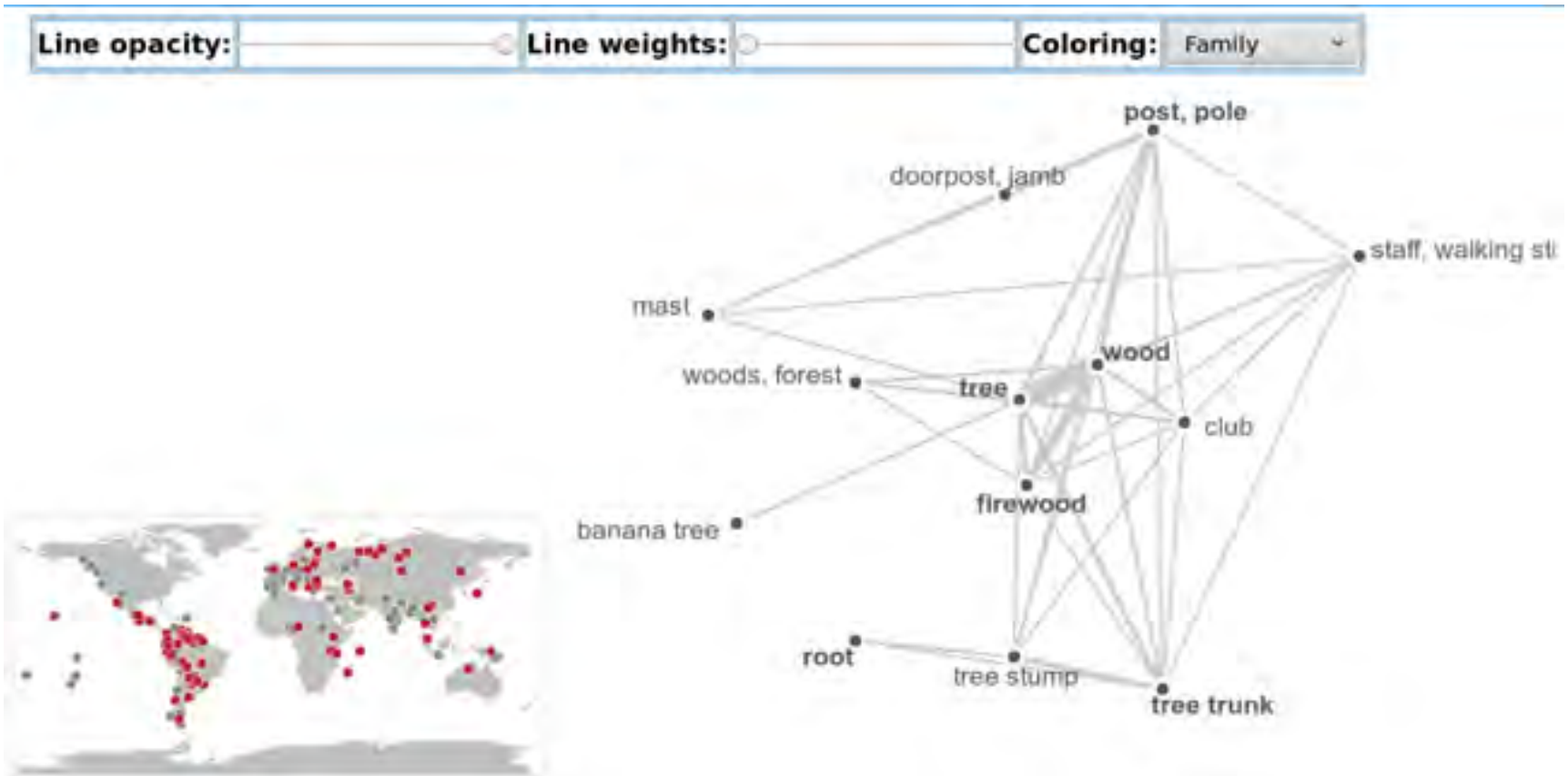
- Lexical-semantic relations:

  TREE – FIREWOOD – FIRE

- Where does 'wood' (e.g. as construction material) come in?

> The vast majority of Sahul languages do not have a distinct lexeme for 'wood'; this is colexified with the lexeme used for English 'tree' . . . In 168 of the Papuan languages surveyed 'tree' and 'wood' were identically colexified (for 37 languages no translation for 'wood' was found in the source used). **In only 11 languages were they recorded as non-identical** and in three of these there appeared to be a possible derivational relationship between the two lexemes. In the Australian languages **only 17 of the 63 languages are recorded as having a word for 'wood' or 'stick' that is different from the word for 'tree'** and in all but six of these there exists a colexified word for 'wood'/'tree' as well. In 11 Australian languages no translation for 'wood' or 'stick' was available. (Schapper et al. 2016: 364/5, our emphasis)

Colexification of ‘wood’ and ‘tree’



'Wood' and 'tree'

- Wiktowski et al. (1981)[10] have claimed that words for 'tree' are originally derived from wood, rather than being generalizations over various tree species.
- Put differently, generic words for 'tree' are primarily associated with the material (wood), rather than 'tree' as a plant species.

[10] S. R. Wiktowski, C. H. Brown, and P. K. Chase (1981). "Where do tree terms come from?" In: *Man* 16.1, pp. 1–14.

Colexifying 'raw material' and 'product'?

- Is there are general tendency for languages (of a specific area) to encode the 'raw material' and the product made from it with the same word?

  "Some – but by no means all – Australian languages take the principle of having a single term to describe some natural object, and also something that can be made from it, to the extreme of having a single lexeme covering both 'tree, wood' and 'fire'." (Dixon 1980: 103)[11]

[11] R. M. Dixon (1980). *The Languages of Australia*. Cambridge: Cambridge University Press.

'Actual' and 'potential' polysemy?

What has been called "actual/potential polysemy" is ubiquitous in Australian languages (cf. Dixon 1980:102-103): a large number of languages exhibit such polysemies as 'cloud/rain'; 'firewood/fire'; 'breast/milk'; 'animal/meat', 'tree/wood/implement', 'bush/bushfire', 'hit/kill', 'sick/ dead', 'dead/rotten'. Significantly, in most Australian languages the word for 'make' is the factitive of 'good' (e.g. Warlpiri *ngurrju* 'good', *ngurrjumani* 'make'). Most of the physical world is regarded, with respect to the transformations that can be worked on it, in the same way that Michelangelo is said to have regarded blocks of marble: as already containing the form of David, etc. Successive phases of potential transformation are then named with the same term.

Evans (1992: 479)[12]

[12] N. Evans (1992). "Multiple semiotic systems, hyperpolysemy, and the reconstruction of semantic change in Australian languages". In: *Diachrony within Synchrony: Language History and Cognition. Papers from the International Symposium at the University of Duisburg, 26–28 March, 1990*. Ed. by G. Kellermann. Frankfurt: Lang, pp. 475–508.

<mountain, stone>

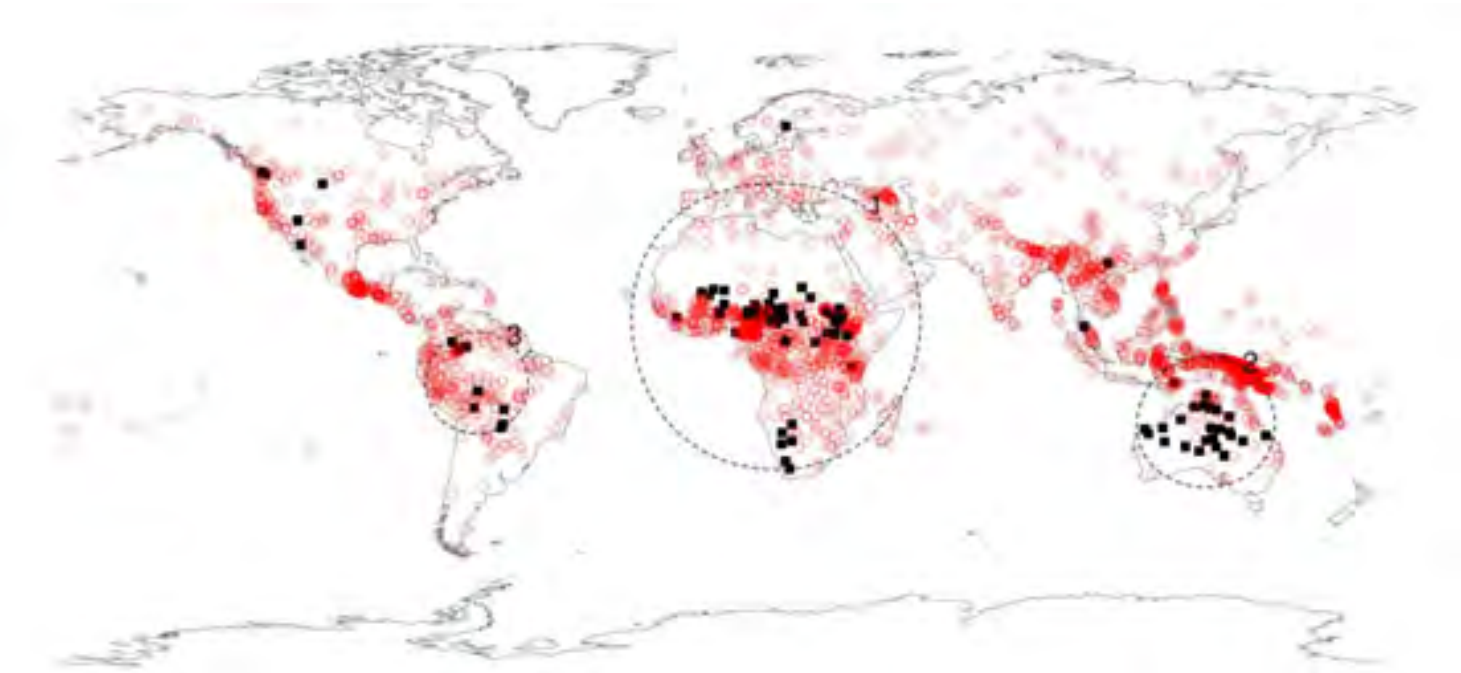
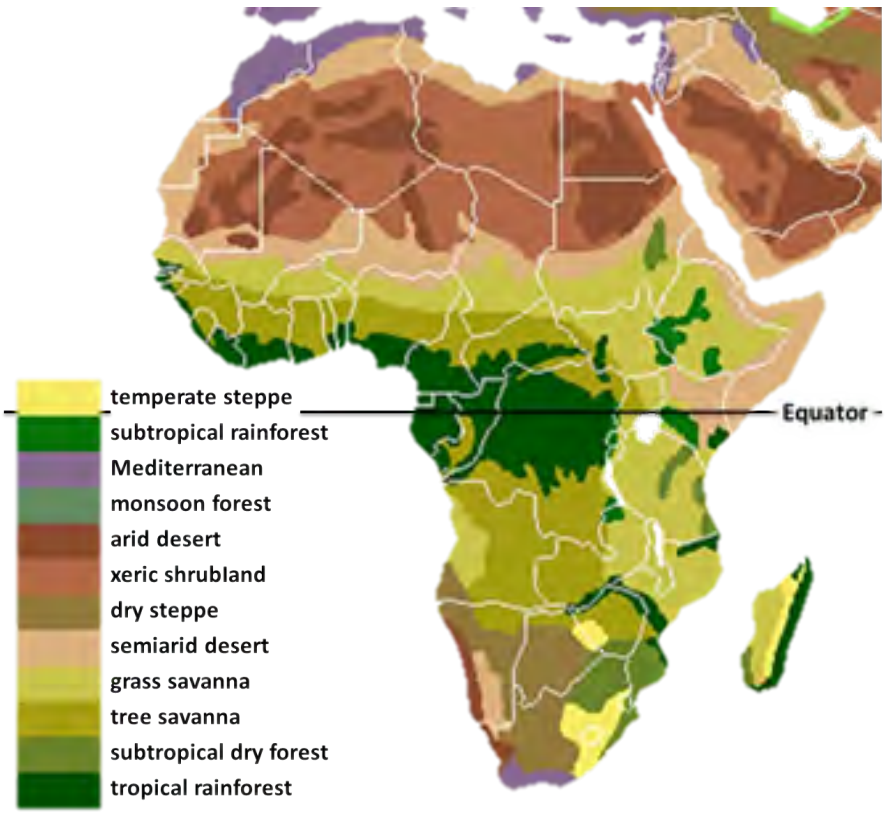


Figure 9: Significant macro-clusters for <'mountain', 'stone'> in the ASJP-data

<mountain, stone> in Australia

Climatic zones of Africa



Aso Rock (near Abuja, the capital of Nigeria)

The Kalahari basin (Khoisan)

Alternative colexification partners for 'mountain'

- A certain affinity between 'mountain' and 'stone' can also be observed in Indo-European languages, e.g. Goth. *hallis* 'rock', Old Norse *hallr* 'large stone', cf. Lat. *collis*; see Buck 1949: Sect. 1.22).[13]
- Alternatively, 'mountain' is often colexified with 'forest', or 'mountain forest', e.g. (Latin American) Spanish *selva* (cf. also Urban 2012).[14]

[13] C. Buck (1949). *A Dictionary of Selected Synonyms in the Principal Indo-European Languages*. Chicago: University of CHicago Press.

[14] M. Urban (2012). "Analyzability and Semantic Associations in Referring Expressions. A Study in Comparative Lexicology". PhD thesis. University of Leiden.

Catamarca (Argentina)

<ear, leaf>

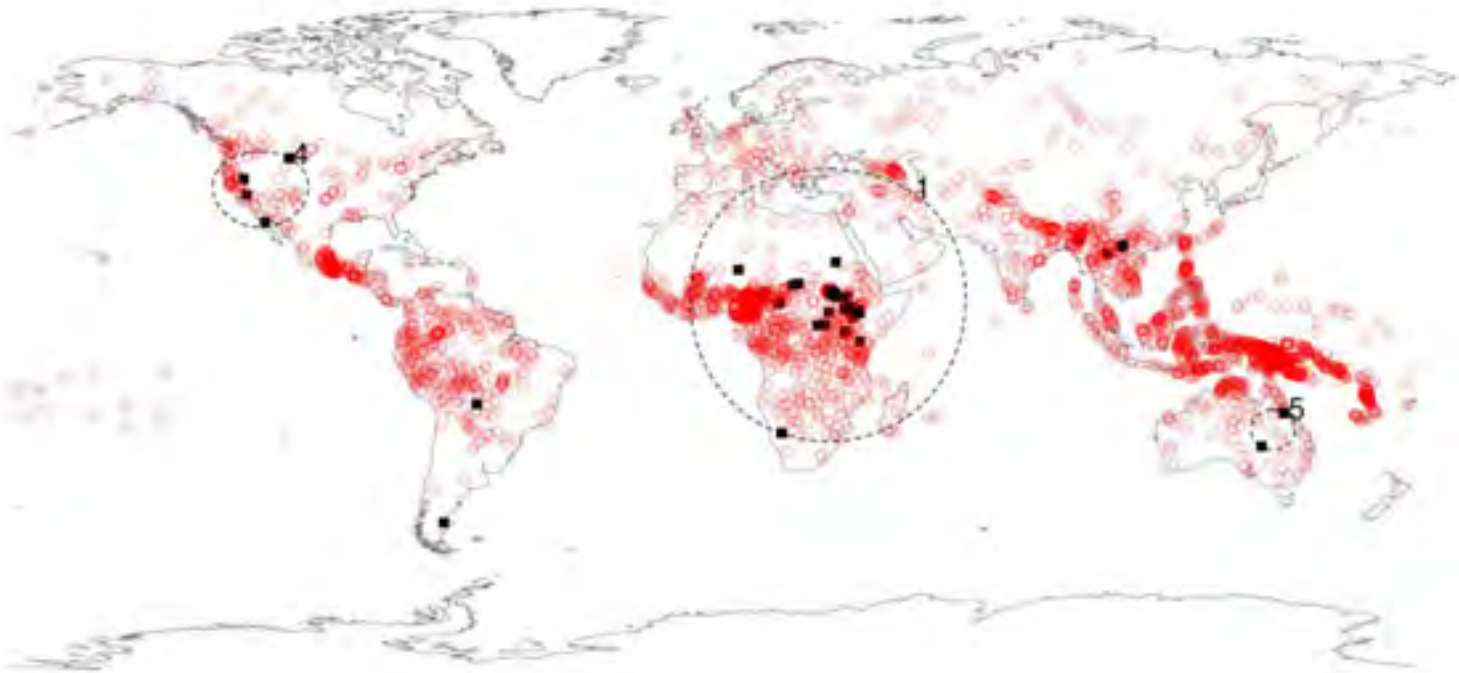


Figure 10: Significant meso-clusters <'ear', 'leaf'> in the ASJP-data

<ear, leaf> in the Nilotic-Surmic spread zone (Nilo-Saharan?)

Map 2: Macro-areas and accretion zones of Afrabia (Güldemann forth. a)

15

[15] T. Güldemann (forthcoming). "Areal linguistics beyond contact and linguistic areas in Afrabia". In: *The Languages and Linguistics of Africa*. Ed. by T. Güldemann. The World of Linguistics. Berlin: de Gruyter Mouton.

<ear, leaf>: Vegetation (ensete/Ethiopian banana)

- Sheko (shek1245): *haay* 'ear', 'leaf of ensete or yam' (Hellenthal 2010: 493)[16]

[16] A.-C. Hellenthal (2010). "A Grammar of Sheko". PhD thesis. Utrecht: LOT.

<ear, leaf>: Vegetation (agave)

<bark, skin>

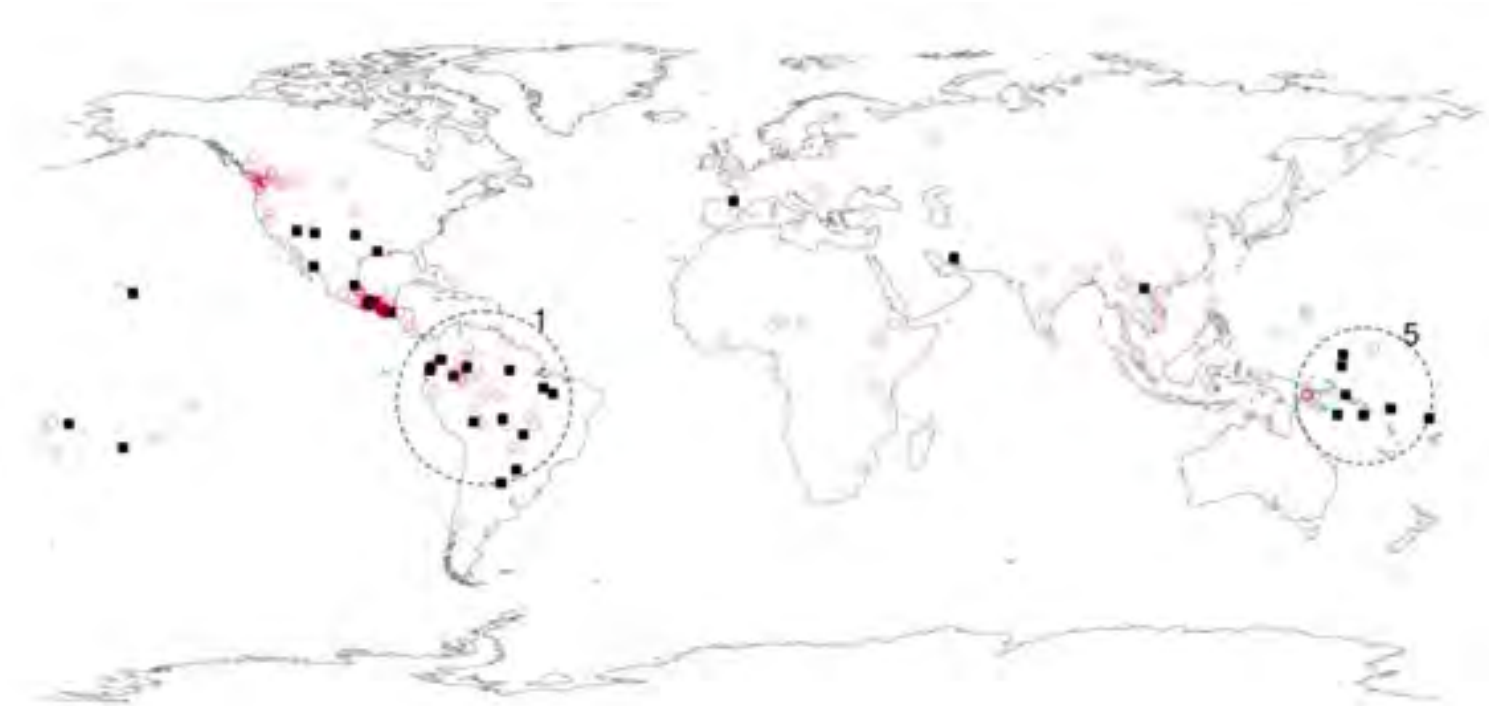


Figure 12: Macro-clusters for <'bark', 'skin'> (ASJP-data)

<bark, skin>: Discussion

- The colexification of 'bark' and 'skin' has been discussed as a potential lexical trait of Mesoamerican languages (Smith-Stark 1994).[17]
- The actual overlap between these concepts is actually more widespread in Mesoamerica, but many of the relevant cases are instances of 'partial colexification'.
- Are such areal patterns reflexes of a more general tendency towards metaphorical transfer between humans, animals and plants?

(1) Copainalá Zoque

ku?yu naka
tree skin

(Harrison et al. 1981: 280)[18]

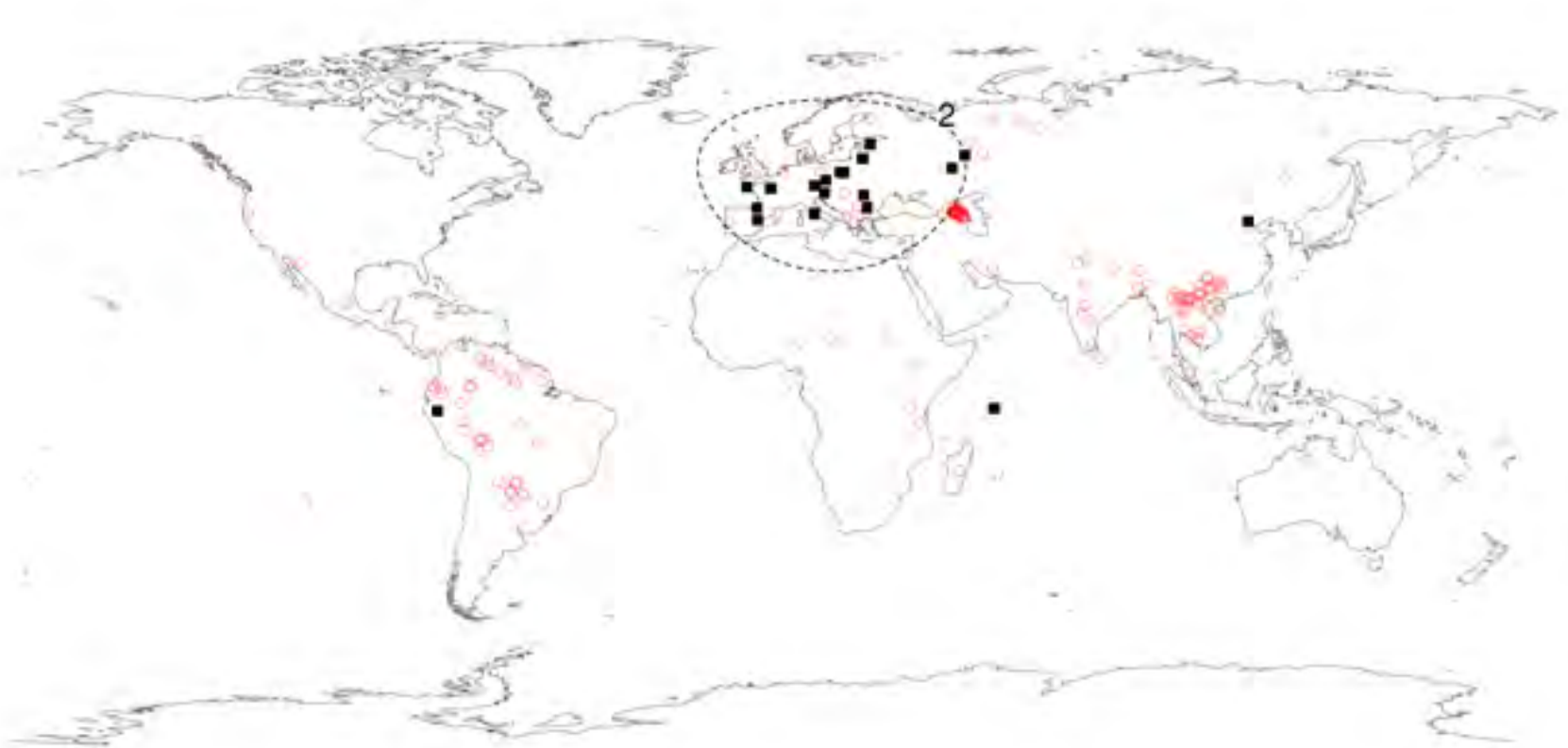
[17] T. C. Smith-Stark (1994). "Mesoamerican calques". In: *Investigaciones lingüisticas en Mesoamérica*. Ed. by C. J. MacKay and V. Vázquez. México City: UNAM, pp. 15–50.

[18] R. Harrison, M. Harrison, and C. Garcia (1981). *Diccionário Zoque de Copainalá*. Mexico: Instituto Linguistico de Verano.

Some results: CLICS

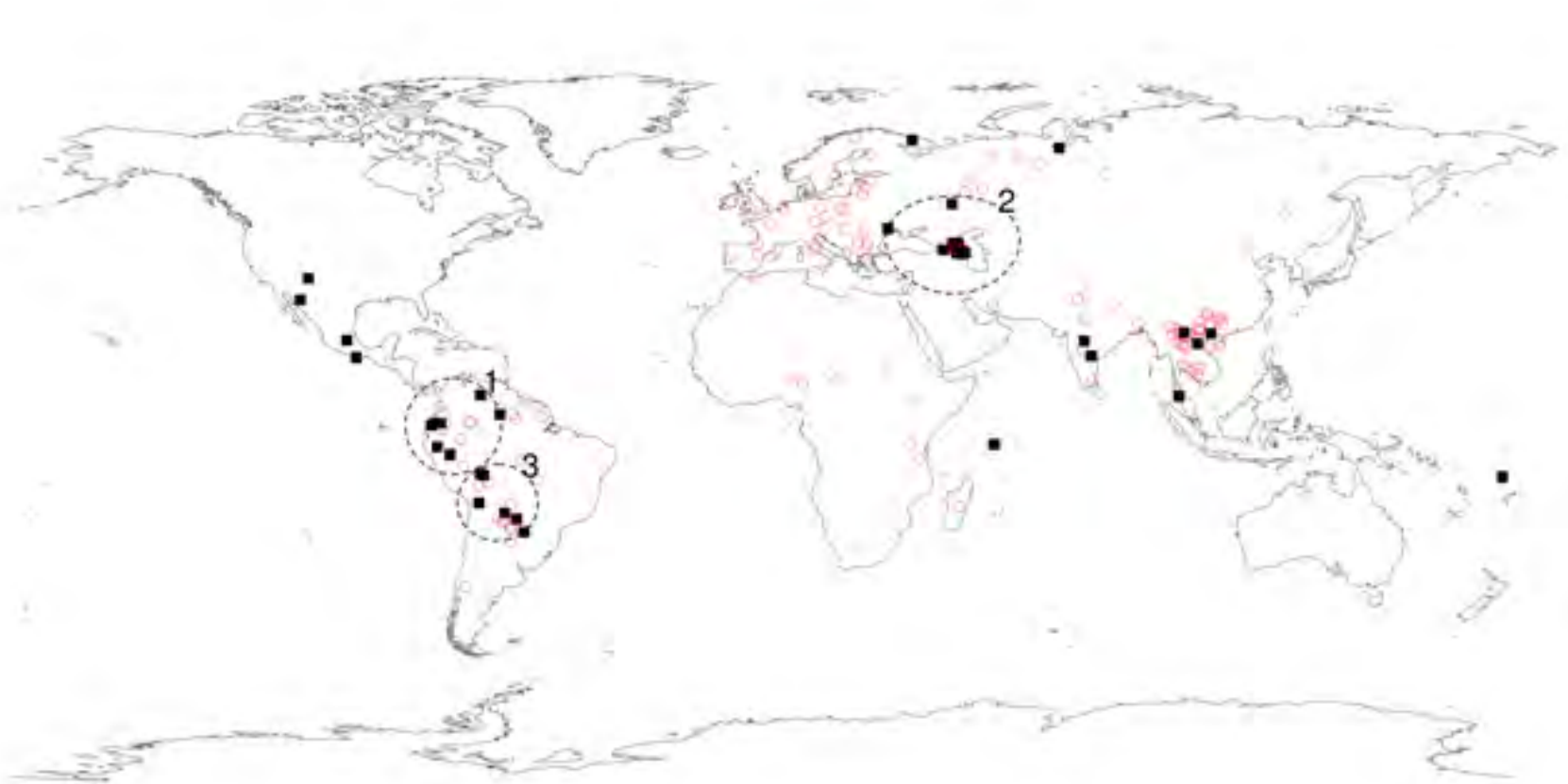
- Some interesting clusters:
  - <feather, pen>
  - <money, x>
  - <language, x>
  - <easy, difficult>

<feather, pen>



<money, silver>

<money, coin>

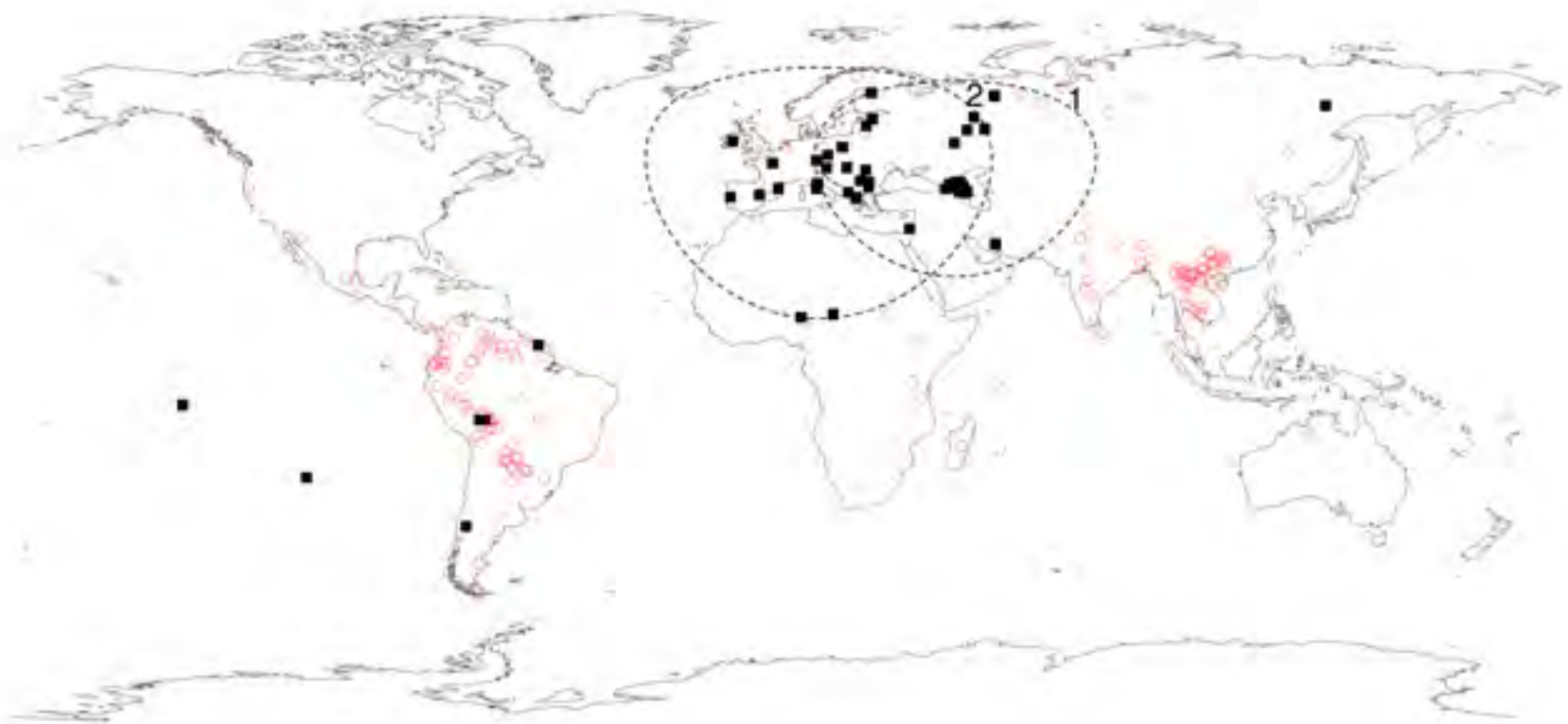


Cultural artifacts: Discussion

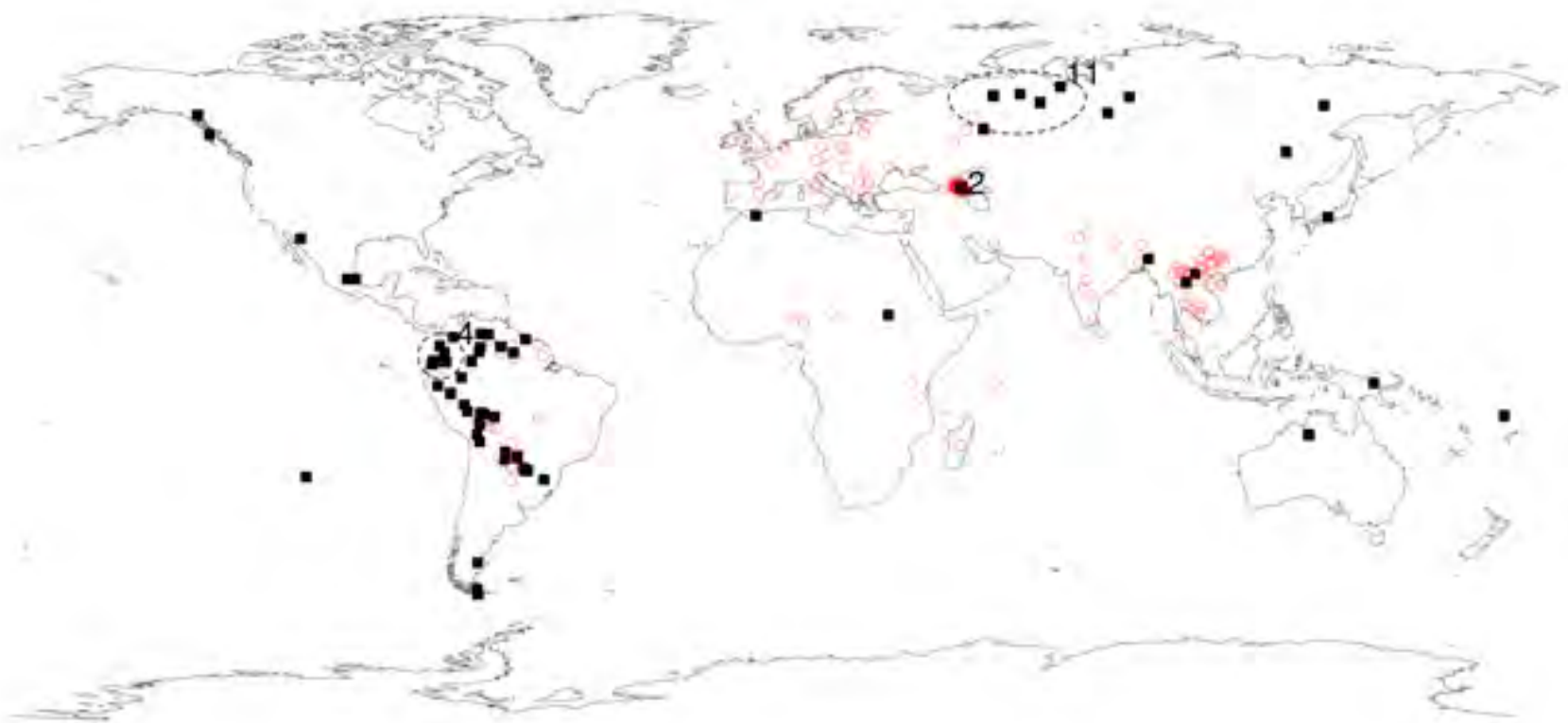
- Terms for cultural artifacts such as 'pen' and 'money' often seem to form relatively well defined clusters of limited size.
- In comparison to other concepts, cultural artifacts are relatively young and were often borrowed or calqued.
- Types of 'lexical acculturation' (Brown 1999):[19]
  - adoption of loan words,
  - borrowing a lexico-constructional pattern,
  - extension of native terms,
  - coining of new expressions.
- Colexification typically seems to result from the extension of native terms ('money' AS 'gold'/'silver', 'shells'/'cowry').
- In some cases colexification may be the result of real-world developments in cultural practices ('pen' AS 'feather', 'money' AS 'silver').

[19] C. Brown (1999). *Lexical Acculturation in Native American Languages*. Oxford: Oxford University Press.

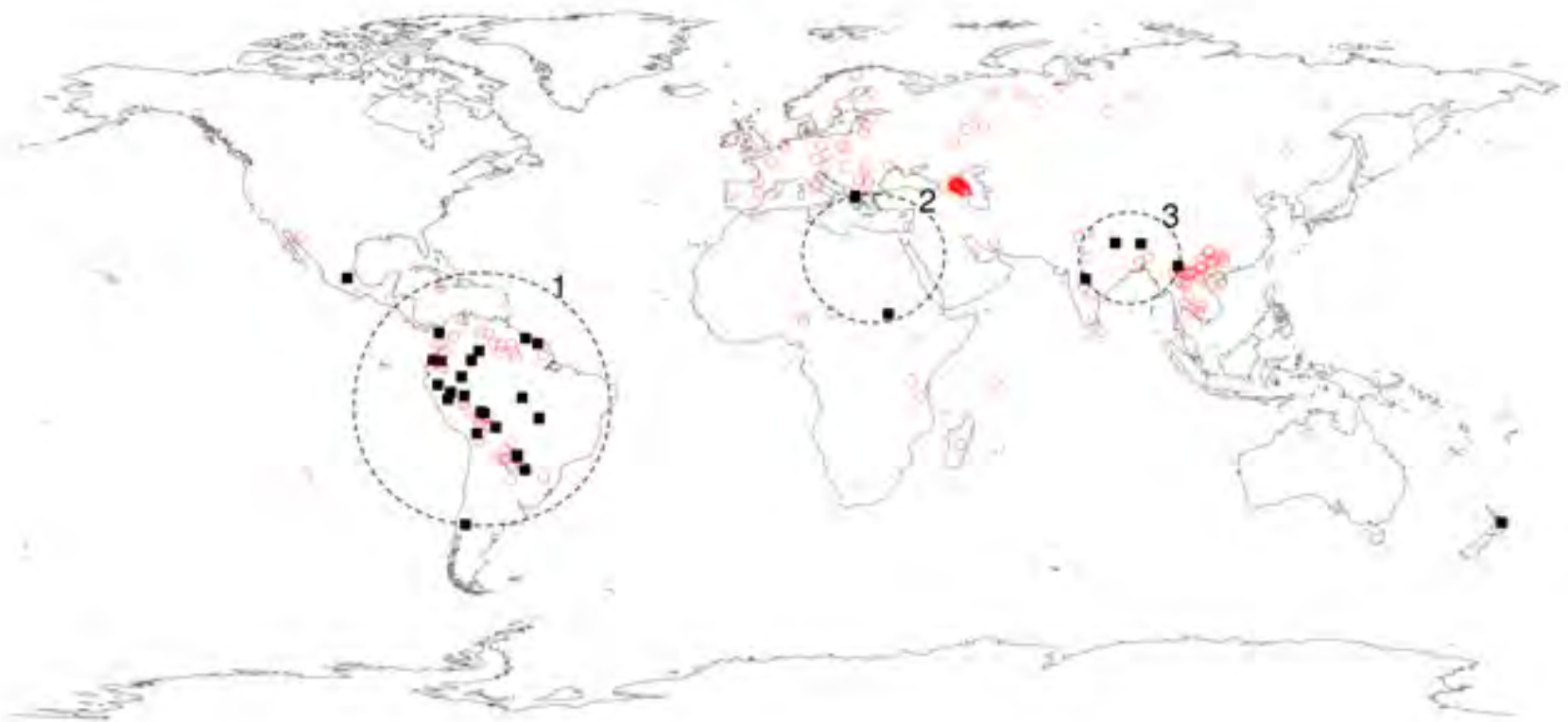
<language, tongue>



<language, word>



<language, voice>



<language, x>: Discussion

- Radden (2004):[20] Terms for language as the result of metonymic change speech organ > speaking > speech > language
- Europe and the Caucasus: instrument-for-action metonymy
- South America: action-for-result metonymy (also specific-for-generic)
- Comparison to cultural artifacts: much larger clusters

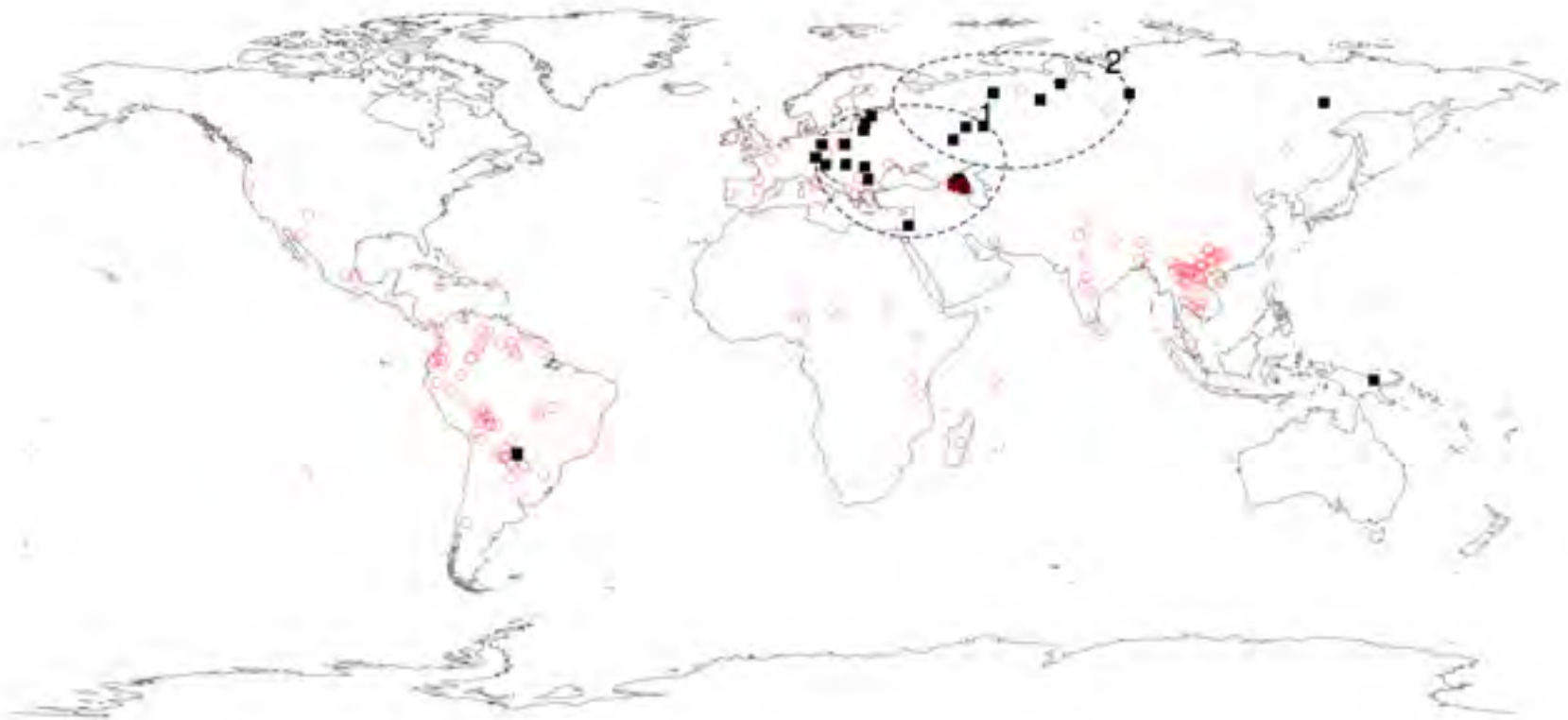
[20] G. Radden (2004). "The metonymic folk model of language". In: *Imagery in language: Festschrift in honour of Professor Ronald W. Langacker*. Ed. by B. Lewandowska.Tomaszczyk and A. Kwiatkowska. Frankfurt: Peter Lang, pp. 543–565.

The concept of 'language'

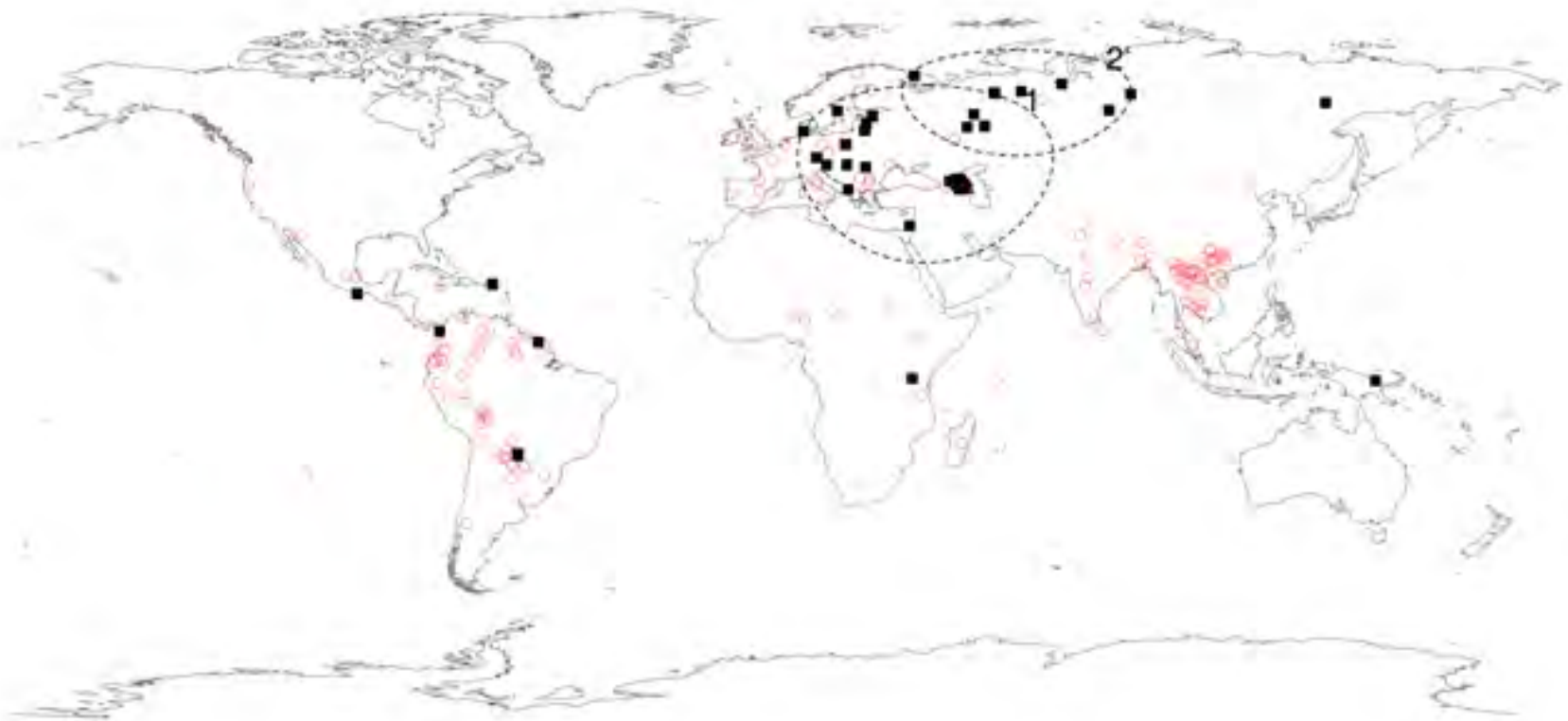
- Goddard (2010):[21] The term 'language' is highly polysemous.
- In expressions such as 'the English language' it has "ideological underpinnings" (Goddard 2010: 44) which are incompatible with certain communities (e.g. of Aboriginal Australia).
- If a community lacks the ideological concept of 'language', the action or the result of speaking ('speech', 'word') will be used as the closest translation equivalent.
- Do areal patterns in colexification reflect such 'ideological' differences?

[21] C. Goddard (2010). "The lexical semantics of language (with special reference to words)". In: *Language Sciences* 33.1, pp. 40–57.

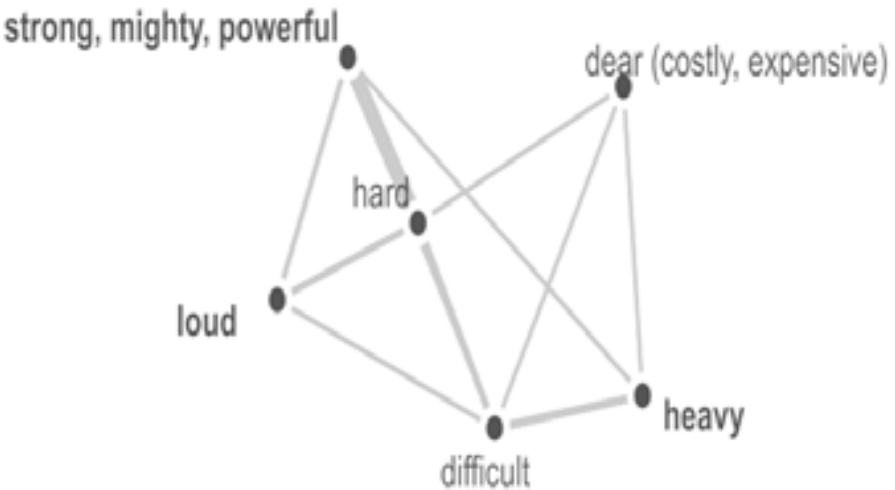
<difficult, heavy>



<easy, light>



Colexification partners of 'difficult' and 'easy' (CLICS)



Colexification of <difficult, heavy>

Found 32 colexifications for "difficult" and "heavy". ?
Note that the number of attested colexifications may differ from the number of languages in which the colexifications were attested.

Nr.	Language	ISO	Family	Source	Form
1	Avar (Zakataly)	ava	North Caucasian	IDS	бакIаб
2	Botlikh	bph	North Caucasian	IDS	гьакIуба
3	Bulgarian	bul	Indo-European	IDS	téžak
4	German, Standard	deu	Indo-European	IDS	schwer
5	Sorbian, Lower	dsb	Indo-European	WOLD	śěžki
6	Estonian	est	Uralic	IDS	raske
7	Hungarian	hun	Uralic	IDS	neheez
8	Judeo-Tat	jdt	Indo-European	IDS	гурунд
9	Khanty	kca	Uralic	IDS	lawaṙt
10	Ket	ket	Yeniseian	WOLD	səə
11	Khvarshi (Khvarshi)	khv	North Caucasian	IDS	луйтIу
12	Kilivila	kij	Austronesian	IDS	гран
13	Karata	kpt	North Caucasian	IDS	гьаркIибоб

Colexification of <difficult, hard>

Found 11 colexifications for "difficult" and "hard". ?
Note that the number of attested colexifications may differ from the number of languages in which the colexifications were attested.

Nr.	Language	ISO	Family	Source	Form
1	Negerhollands	dcr	Creole	IDS	tai
2	Emberá, Northern	emp	Choco	IDS	za'rea
3	Iraqw	irk	Afro-Asiatic	WOLD	gawid
4	Marathi	mar	Indo-European	SPRÅKBANKEN	kaʈʰi:ɳ
5	Macushi	mbc	Carib	IDS	saʔme
6	Quichua, Imbabura Highland	qvi	Quechuan	WOLD	sinchi
7	Tarifit	rif	Afro-Asiatic	WOLD	qsəħ
8	Saramaccan	srm	Creole	WOLD	taánga
9	Sirionó	srq	Tupi	IDS	eãtã
10	Swahili	swh	Niger-Congo	WOLD	gumu
11	Waurá	wau	Arawakan	IDS	kehežuka

<difficult, easy>: More data

- The <difficult, heavy> colexification is also found in the Middle East and (North-)East Africa, e.g. (cf. Segerer and Flavier 2011–2018):[22]
  - Hebrew *kaved*/*kveda*
  - Arabic *ṯaqīl*
  - Amharic *käbbad*
  - Sheko *ints*
- Some languages of West Africa (Senegal) seem to have it, too, e.g.:
  - Soninke (Mande)
  - Joola-Manjaku, Joola-Fonyi (Atlantic-Congo)

[22] G. Segerer and S. Flavier (2011–2018). *RefLex: Reference Lexicon of Africa. Version 1.1.* http://reflex.cnrs.fr. Paris, Lyon.

<difficult, easy>: Discussion

- The metaphorical extension of 'heavy' to 'difficult' is regarded as a 'primary metaphor' in cognitive linguistics, as is the extension from 'hard' to 'difficult' (assumed universality, cf. Grady 2005).[23]
- But then, the distribution of <hard, difficult> seems to exhibit heavy areal (as well as genealogical) biases.
- Hypothesis: There are various (universal) options of conceptualizing 'difficulty', and there is no obvious motivation for preferring any of them (e.g. 'heavy' vs. 'hard').
- Note that metaphorical extensions of this type may, to some extent, be externally motivated, cf. 'warm' as a positive emotional feeling.
- Does the (cultural) 'arbitraniness' of metaphorial transfer of this type give language contact more weight?

[23] J. Grady (2005). "Primary metaphors as inputs to conceptual integration". In: *Journal of Pragmatics* 37, pp. 1595–1614.

Limitations of our approach

- Adjectives such as 'difficult' are complex semantically speaking and often imply metonymic shifts, e.g.:
  - This is a difficult task.
  - He's a difficult person.
  - Those are difficult times.
  - etc.
- Some words, such as Russ. *tâželyj*, can only be used with the meaning 'difficult' in specific contexts.

'Difficult' and 'easy' in German and Czech

- A corpus-based comparison of Germ. *schwer*, *schwierig* and Cz. *těžký*, *obtížný* (198 examples from the Europarl corpus).
- The adjectives of Czech are distributed differently from those of German, insofar as *těžký* shows a clear affinity to temporal head nouns (in the data investigated).

(2) . . . I think we have a person who has already shown in those very **difficult days** in Pakistan how responsibly he has been taking on these things . . .

(3) . . . máme myslím člověka, který již během těchto velmi **těžkých dní** v Pákistánu ukázal, jak zodpovědně k těmto věcem přistupuje . . .

Convergence areas and types of lexical meaning

- What motivates colexification patterns in an areal perspective?
- What general tendencies can we observe with respect to types of lexical meanings and convergence areas (also in terms of size)?
- Cultural artifacts seem to form relatively clear-cut areal clusters, perhaps because of a relatively shallow time depth.
- Some colexification patterns seem to be related to (physical) properties of the environment.
- Are there general (areal) tendency towards specific types of metaphorical transfer, and perhaps "actual/potential polysemy" (Evans 1992)?
- Abstract concepts are sometimes associated with larger cluster areas.
- Problem: Words as elicitation stimuli/*tertia comparationis* may actually be misleading ('language').

Thanks for listening!

- Any suggestions?