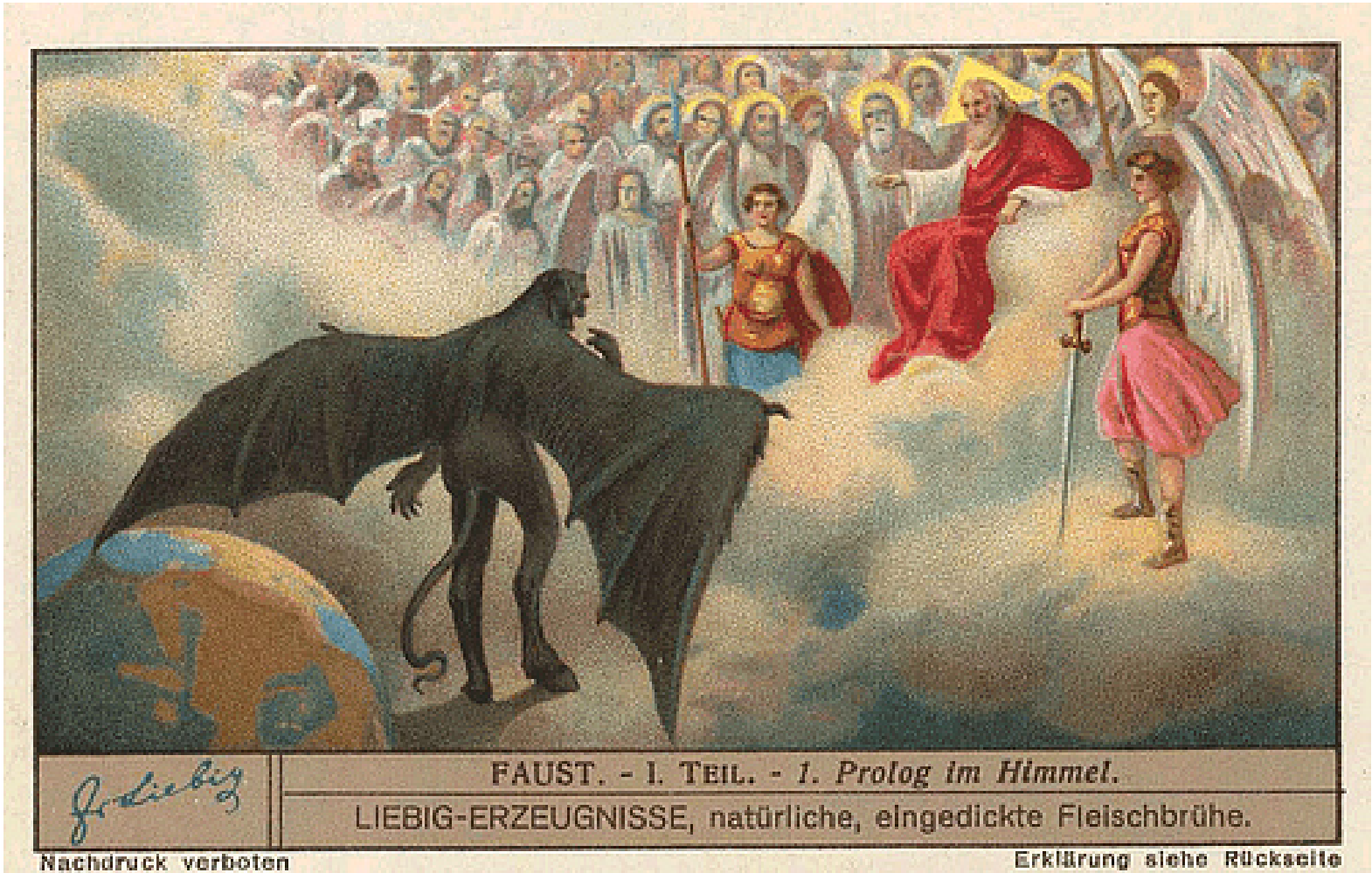
Concepticon: A Resource for the Linking of Concept Lists

Johann-Mattis List[1], Michael Cysouw[2], and Robert Forkel[3]

[1]CRLAO/EHESS and AIRE/UPMC, Paris
[2]Forschungszentrum Deutscher Sprachatlas, Marburg
[3]Max Planck Institute for Evolutionary Anthropology, Leipzig

2015-04-30


FAUST. - I. Teil. - 1. Prolog im Himmel.
LIEBIG-ERZEUGNISSE, natürliche, eingedickte Fleischbrühe.
Nachdruck verboten
Erklärung siehe Rückseite

Prologue

[...] it is a well known fact that certain types of morphemes are relatively stable. Pronouns and numerals, for example, are occasionally replaced either by other forms from the same language or by borrowed elements, but such replacement is rare. The same is more or less true of other everyday expressions connected with concepts and experiences common to all human groups or to the groups living in a given part of the world during a given epoch.

Morris Swadesh (1950): Salish internal relationships. *IJAL* 16.4.

[...] it is a well known fact that certain types of morphemes are relatively stable. Pronouns and numerals, for example, are occasionally replaced either by other forms from the same language or by borrowed elements, but such replacement is rare. The same is more or less true of other everyday expressions connected with concepts and experiences common to all human groups or to the groups living in a given part of the world during a given epoch.

Morris Swadesh (1950): Salish internal relationships. *IJAL* 16.4.

[...] it is a well known fact that certain types of morphemes are relatively stable. Pronouns and numerals, for example, are occasionally replaced either by other forms from the same language or by borrowed elements, but such replacement is rare. The same is more or less true of other everyday expressions connected with concepts and experiences common to all human groups or to the groups living in a given part of the world during a given epoch.

I, thou, he, we, ye, one, two, three, four, five, six, seven, eight, nine, ten, hundred, all, animal, ashes, back, bad, bark, belly, big, black, blood, bone, brother (elder), child (son or daughter), cloud, cold, come, cry (weep), dance, day, dog, dust, ear, earth, eat, egg, eye, far, father, fire, flower, fog, foot, good, grass, green, hair, hand, head, heart, here, hit (with fist), hunt, husband, ice, lake, laugh, leaf, left hand, leg, liver, long, louse, man, meat, mother, mountain, mouth, name, near, neck, night, nose, person, rain, red, right hand, road (trail), root, rope, salt, sand, short, sing, sister (elder), skin, sky, small, smoke, snake, snow, speak, spear (war), star, stone, sun, swim, tail, that, there, this, tongue, tooth, tree, warm, water, what, where, white, who, wife, wind, woman, year, yellow.

Morris Swadesh (1950): Salish internal relationships. *IJAL* 16.4.

[...] it is a well known fact that certain types of morphemes are relatively stable. Pronouns and numerals, for example, are occasionally replaced either by other forms from the same language or by borrowed elements, but such replacement is rare. The same is more or less true of other [illegible] day expressions connected with concepts and experiences common to all human groups or to the groups living in a given part of the world during a given epoch.

I, thou, he, we, ye, one, two, three, four, five, six, seven, eight, nine, ten, hundred, all, animal, ashes, back, bad, bark, belly, big, black, blood, bone, brother (elder), child (son or daughter), cloud, cold, come, cry (weep), dance, day, dog, dust, ear, earth, eat, egg, [illegible]r, father, fire, flower, fog, foot, [illegible]s, green, hair, hand, head, heart, [illegible]nit (with fist), hunt, husband, ice, [illegible]e, laugh, leaf, left hand, leg, liver, long, louse, man, meat, mother, mountain, mouth, name, near, neck, night, nose, person, rain, red, right hand, road (trail), root, rope, salt, sand, short, sing, sister (elder), skin, sky, small, smoke, snake, snow, speak, spear (war), star, stone, sun, swim, tail, that, there, this, tongue, tooth, tree, warm, water, what, where, white, who, wife, wind, woman, year, yellow.

Swadesh List

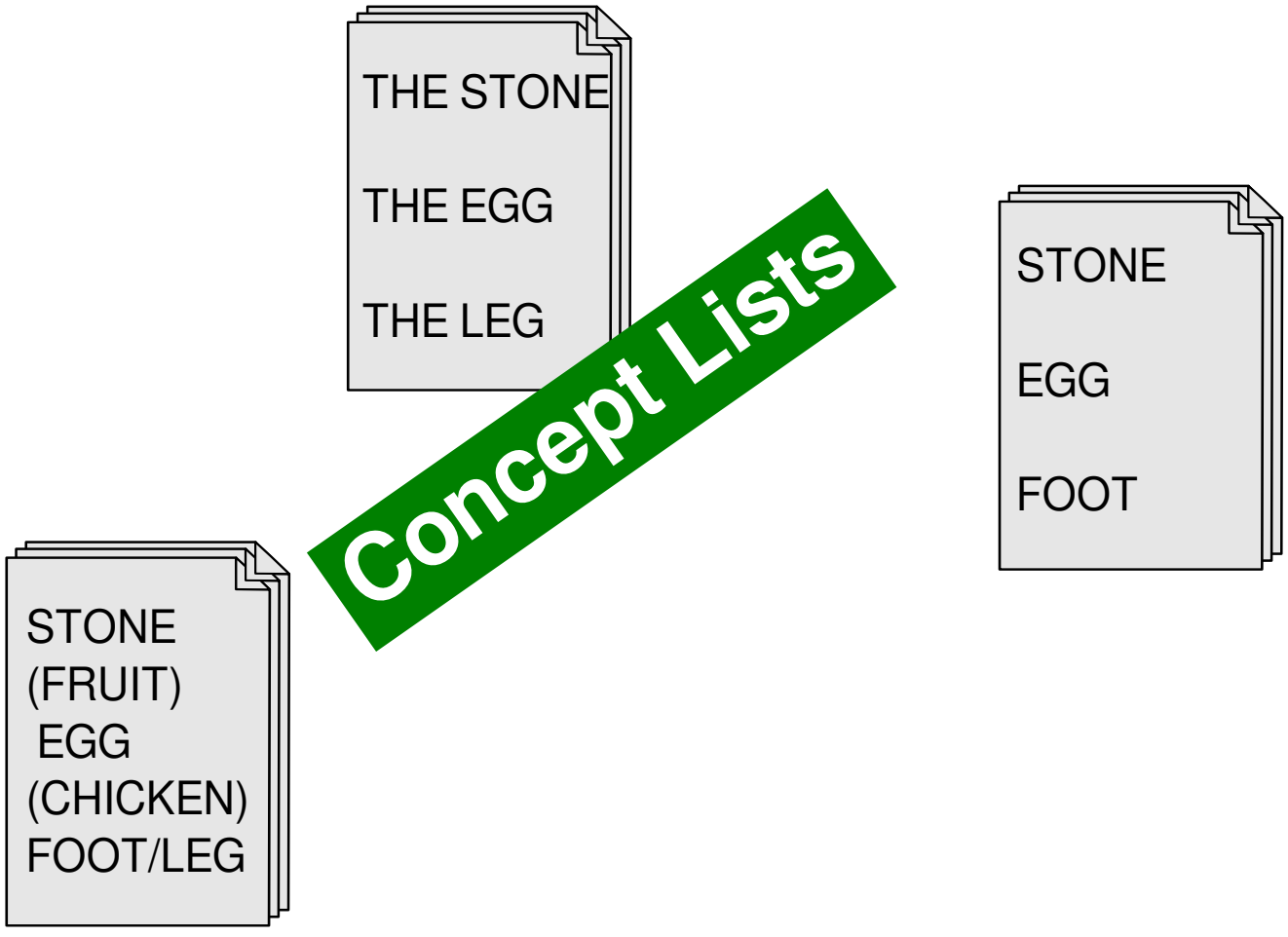
Morris Swadesh (1950): Salish internal relationships. *IJAL* 16.4.

[...] it is a well known fact that certain types of morphemes are relatively stable. Pronouns and numerals, for example, are occasionally replaced either by other forms from the same language or by borrowed elements, but such replacement is rare. The sam more or less true of othe day expressions connec with concepts and experiences common to all human groups or to the groups living in a given part of the world during a given epoch.

Concept Lists

I, thou, he, we, ye, one, two, three, four, five, six, seven, eight, nine, ten, hundred, all, animal, ashes, back, bad, bark, belly, big, black, blood, bone, brother (elder), child (son or daughter), cloud, cold, come, cry (weep), d nce, day, dog, dust, ear, earth, eat, egg, , father, fire, flower, fog, foot, g , green, hair, hand, head, heart, t (with fist), hunt, husband, ice, e, laugh, leaf, left hand, leg, liver, long, louse, man, meat, mother, mountain, mouth, name, near, neck, night, nose, person, rain, red, right hand, road (trail), root, rope, salt, sand, short, sing, sister (elder), skin, sky, small, smoke, snake, snow, speak, spear (war), star, stone, sun, swim, tail, that, there, this, tongue, tooth, tree, warm, water, what, where, white, who, wife, wind, woman, year, yellow.

Morris Swadesh (1950): Salish internal relationships. *IJAL* 16.4.


THE STONE
THE EGG
THE LEG
Concept Lists
STONE
EGG
FOOT
STONE
(FRUIT)
EGG
(CHICKEN)
FOOT/LEG

What are Concept Lists?

Simply speaking, concept lists are lists of concepts, in which concepts are ideally given by both glosses and short definitions. They can be compiled for different purposes (language comparison, concept comparison) and be expanded by adding structure (rankings, divisions, relations).

What is their Purpose?

Language Comparison (historical linguistics, dialectology)

- proving genetic relationship (Yakhontov 1991/35 items, Dolgopolsky 1964/15 items)
- linguistic subgrouping (Norman 2003/40 items, Swadesh 1955/100 items, Starostin 1991/110 items)
- layer identification (Chén 1996/100+100 items, Yakhontov 1991/35+65 items)

Concept Comparison (historical linguistics, psycholinguistics)

- synchronic (word association: SimLex, Hill et al. 2014/1028 items, colexification: CLICS, List et al. 2014/1280 items)
- diachronic (semantic shift: DatSemShift, Bulakh et al. 2013/2424 items, stability of form-meaning relations: WOLD, Haspelmath & Tadmor 2009/1460 items)

What is their Structure?

Type	Example	Purpose
basic vocabulary list ("Swadesh list")	Swadesh 1952 / 200 items	subgrouping
subdivided concept list	Yakhontov 1991 / 35 + 65 items	genetic relationship, layer identification
"ultra-stable" concept list	Dolgopolsky 1964 / 15 items	genetic relationship
questionnaire	Allen 2007 / 500 items	dialect / language comparison
ranked list	Starostin 2007 / 110 items	subgrouping, layer identification
list of concept relations	DatSemShift, Bulakh et al. 2013 / 2424 items	representation of concept relations
special-purpose concept list	Matisoff 1978 / 200 items	subgrouping of Tibeto-Burman languages
historical concept list	Leibniz 1768 / 128 items	language comparison

Examples

NUMBER	RUSSIAN	ENGLISH
1	кровь	blood
2	кость	bone
3	умереть	die
4	собака	dog
5	ухо	ear
6	яйцо	egg
7	глаз	eye
8	огонь	fire
...	...	...

Jakhontov 1991 / 35 items

Examples

NUMBER	ENGLISH
1	belly (exterior)
2	blood
3	bone
4	ear/hear
5	egg
...	...
200	drive/hunt
200a	burn
200b	cut

Matisoff 1978 / 200 items

Examples

NUMBER	CHINESE	GLOSS
1	我	I
2	你	you
3	我们	we
4	这	this
5	那	that
..	..	..
92	晚上	night
93	热	hot
...	...	...

Chén 1996 / 100 items (stable sublist)

Examples

NUMBER	LATIN	CATEGORY	GLOSS
1	unum	Nomina numeralia	one
...	...	...	...
19	avus	Propinquitates & aetates	grandfather
...	...	...	...
35	caro	Partes corporis	flesh
...	...	...	...
82	deus	Naturalia	god
...	...	...	...
128	velle	Actiones	want

Leibniz 1768 / 128 items

Concepts

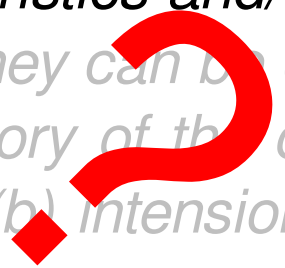
What are Concepts?

Idea which is conceived through abstraction and through which objects or states of affairs are classified on the basis of particular characteristics and/or relations. Notions are represented by terms. They can be defined like sets: (a) extensionally, by an inventory of the objects that fall under a particular concept; and (b) intensionally, ...

(Bussmann 1996: 815)

What are Concepts?

Idea which is conceived through abstraction and through which objects or states of affairs are classified on the basis of particular characteristics and/or relations. Notions are represented by terms. They can be defined like sets: (a) extensionally, by an inventory of the objects that fall under a particular concept; and (b) intensionally, ...



(Bussmann 1996: 815)

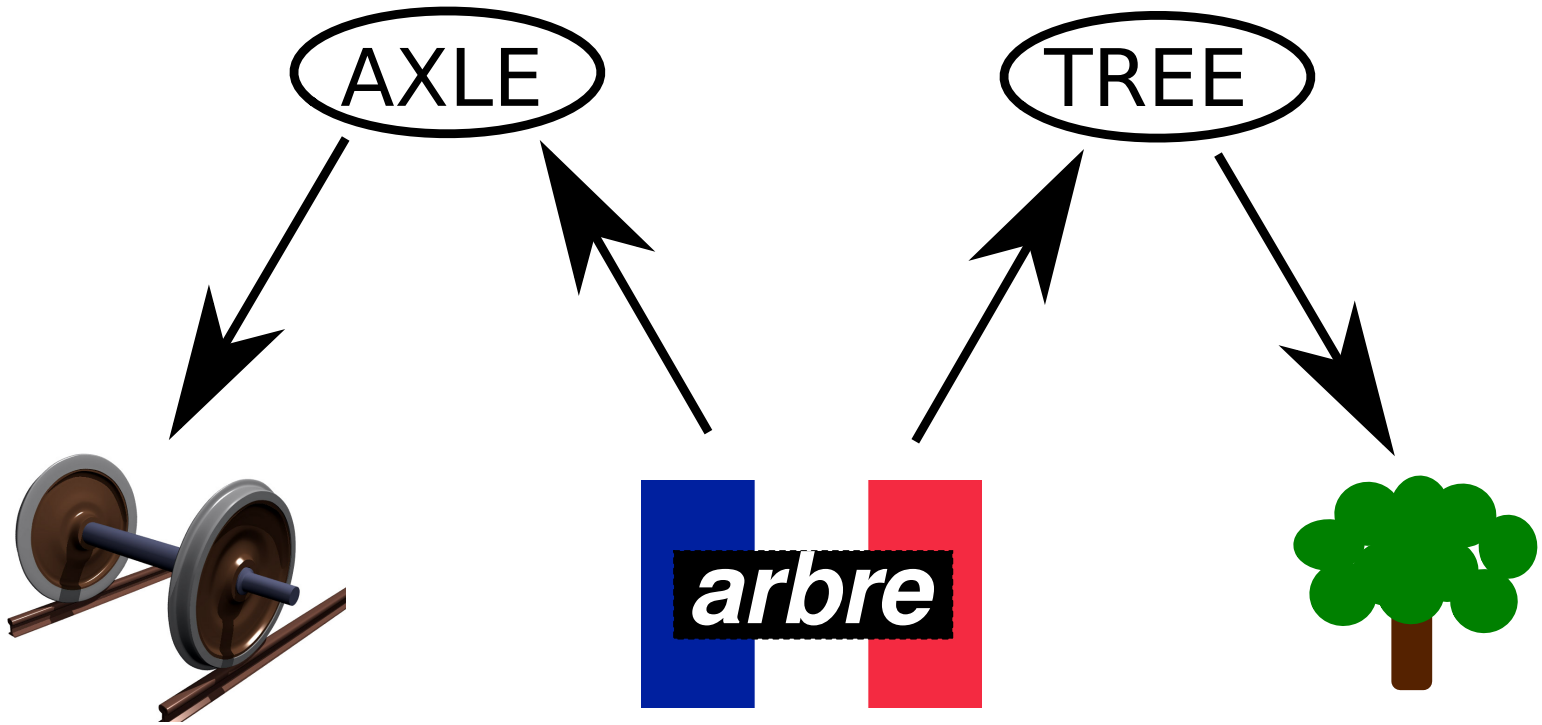
What are Concepts?

Concepts are well-defined objects in semantic space. In historical linguistics, we refer to them with help of English glosses or small definitions. (Very & Simple to appear)

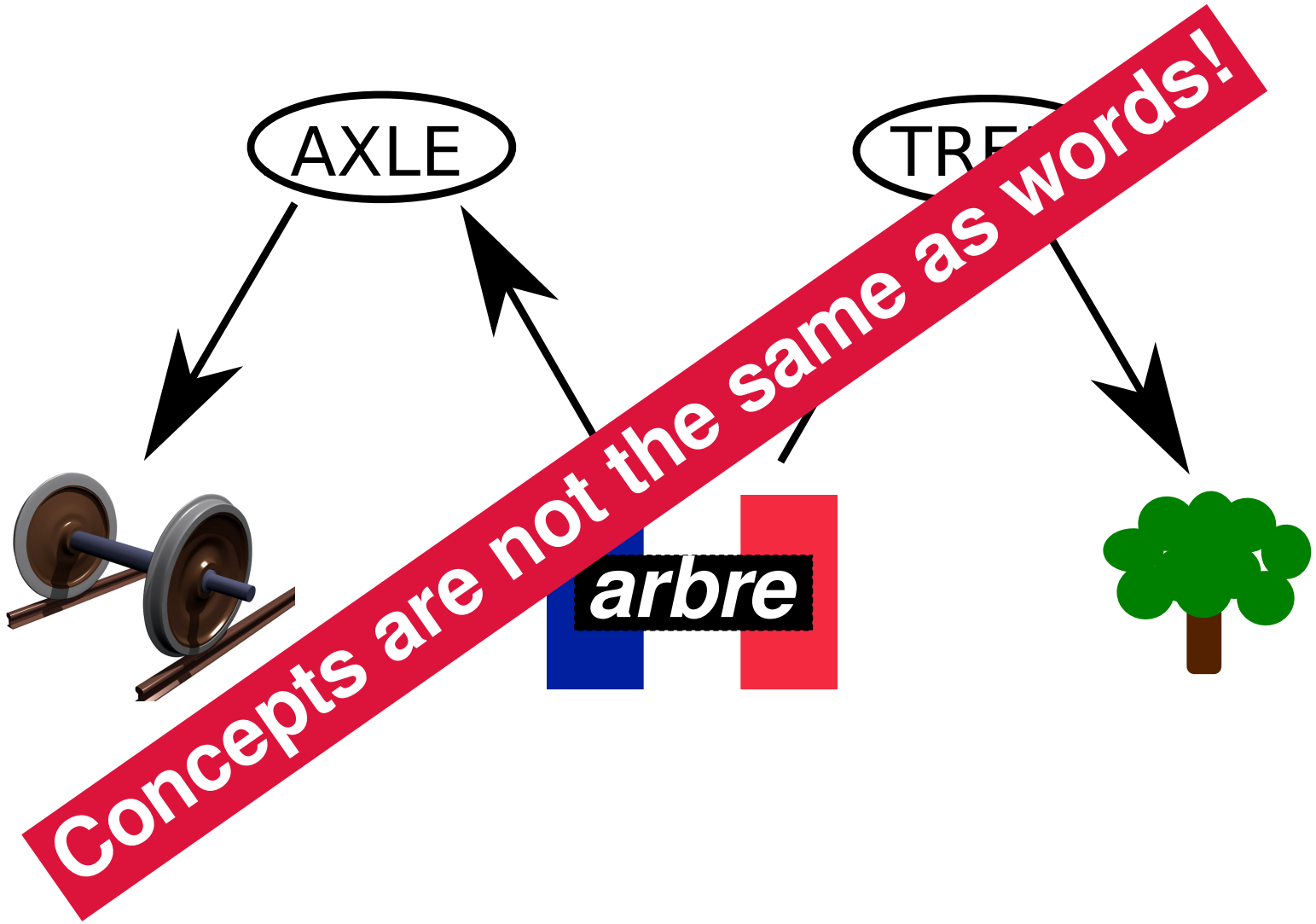
“dog”

“A common four-legged animal, especially kept by people as a pet or to hunt or guard things.”

What are Concepts?



What are Concepts?



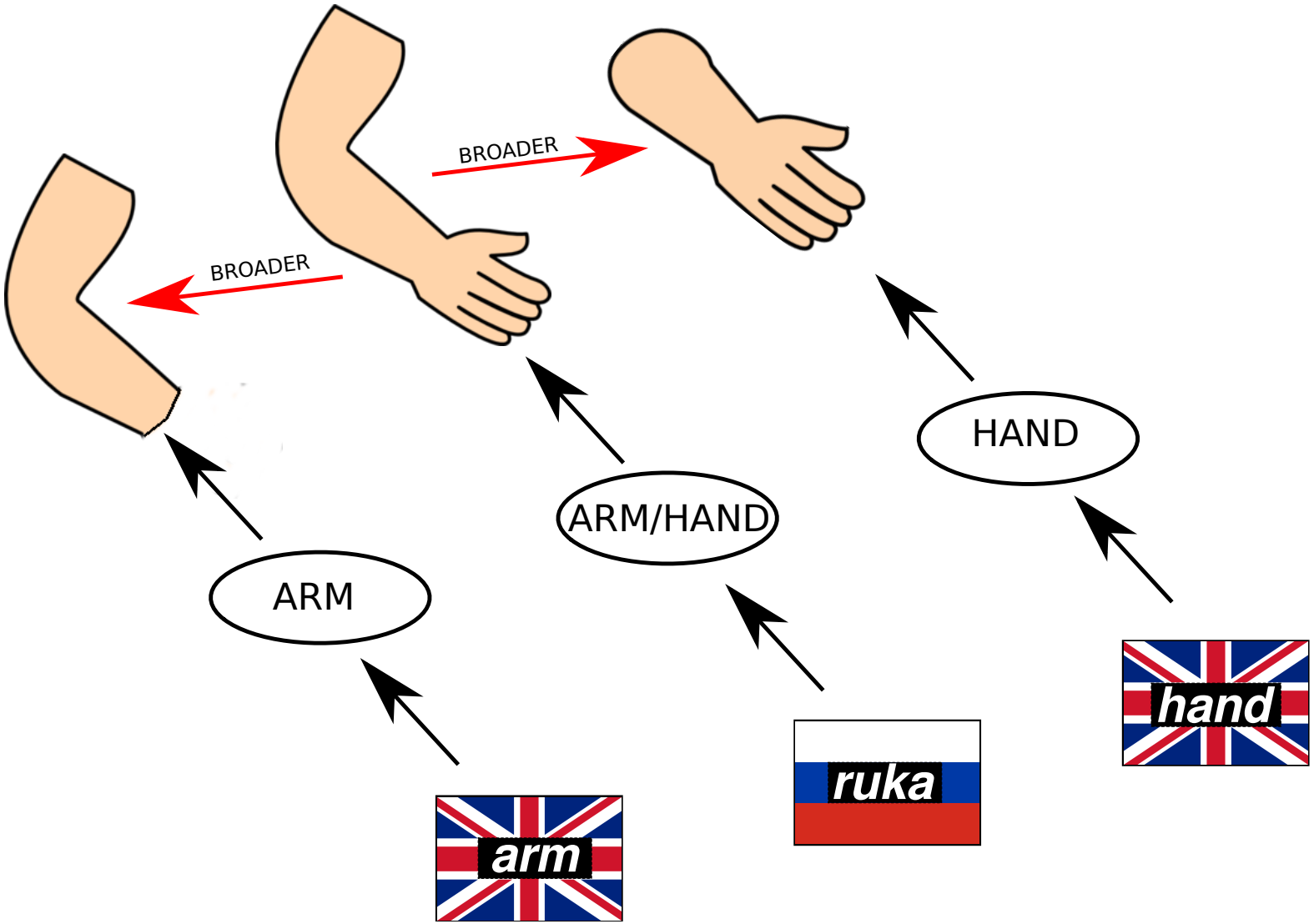
Relations among Concepts

When defining concepts as well-defined objects in some semantic space, it is clear that different relations can be postulated for different concepts.

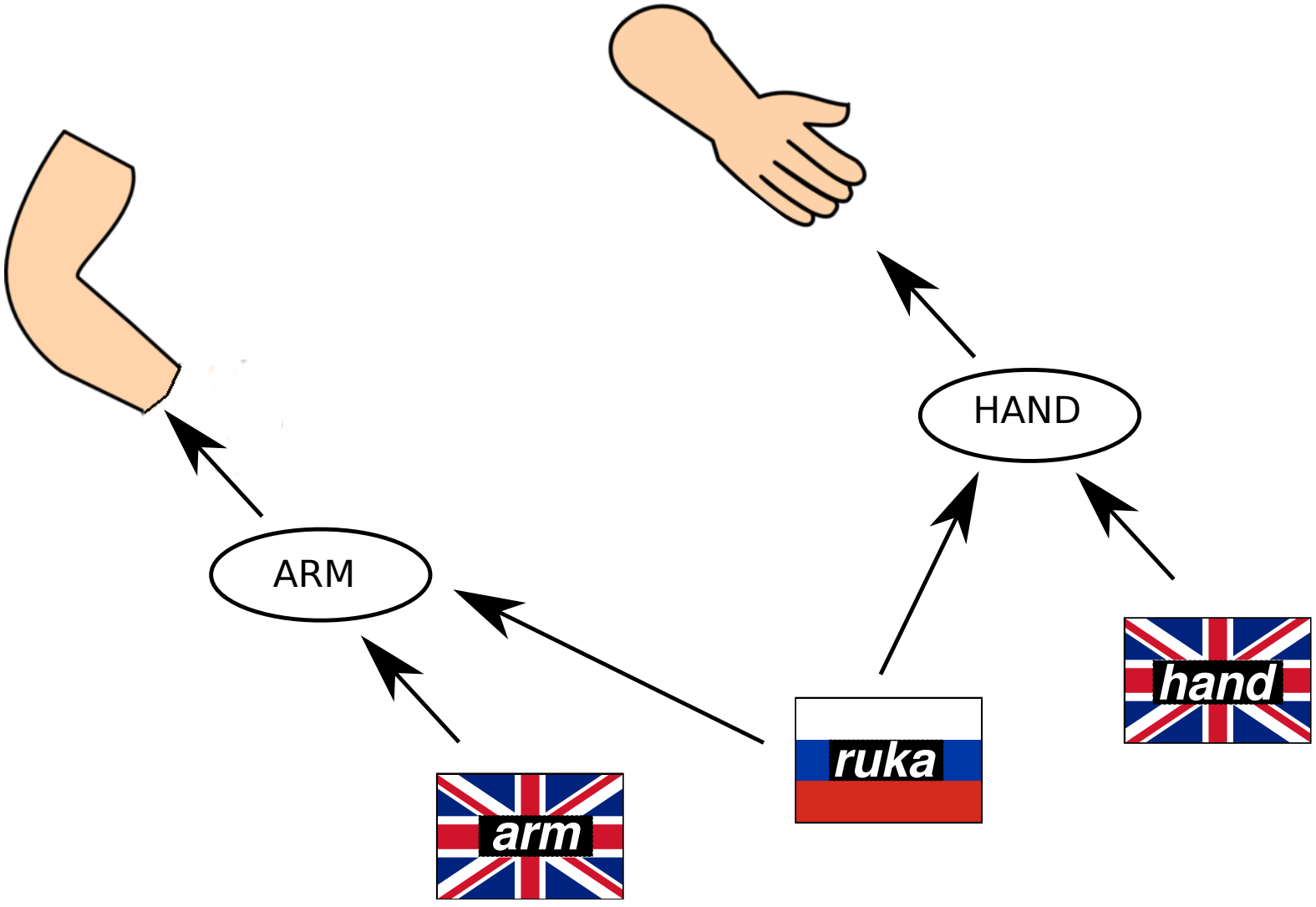
“uncle” is broader than “paternal uncle”

“uncle” is narrower than “one’s parents’ brother or sister”

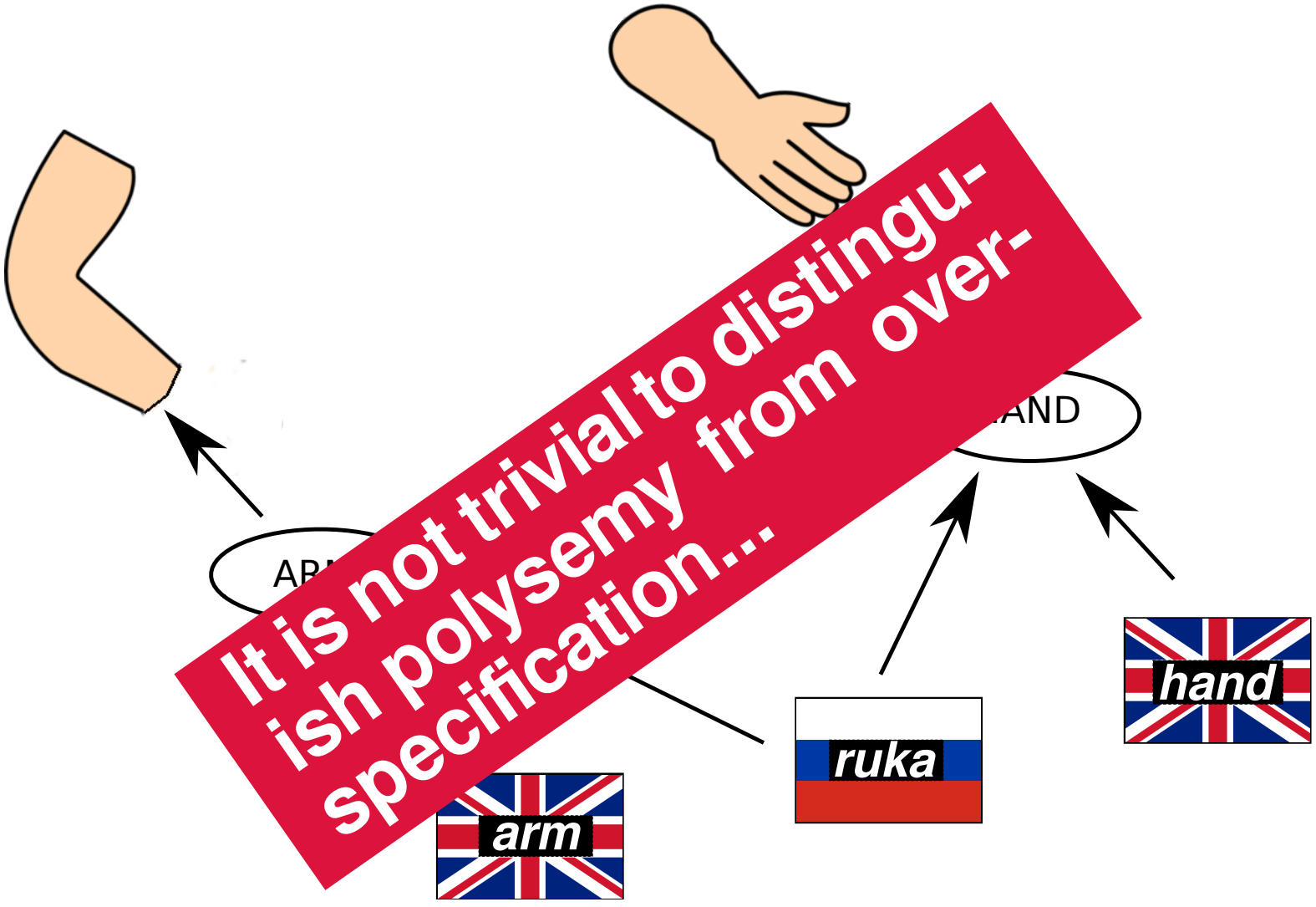
Relations among Concepts

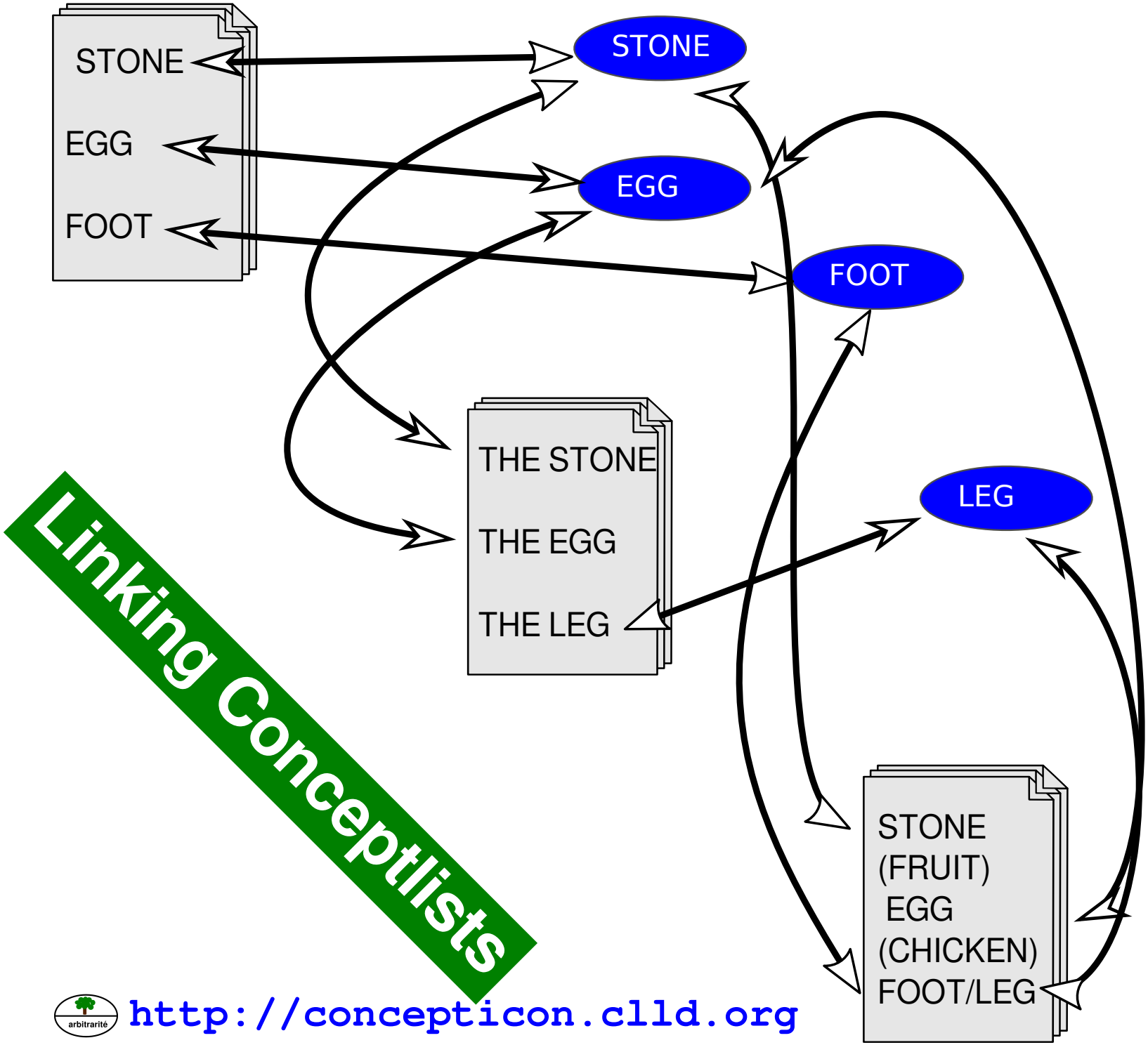


Relations among Concepts



Relations among Concepts




STONE
EGG
FOOT
STONE
EGG
FOOT
THE STONE
THE EGG
THE LEG
LEG
STONE
(FRUIT)
EGG
(CHICKEN)
FOOT/LEG
Linking Conceptlists
arbitrarité
http://concepticon.clld.org

Why to Link Concept lists?

Did you use a specific Swadesh list for your study?

Sure. What would you think?

Which one did you use?

Aren't they all the same?

No...

Really? Well, some from the internet, but it's not important anyway, since I changed a few items. It was too difficult to translate some concepts into my languages...

Why to Link Concept lists?

- facilitating the combination of different datasets
- facilitating the enrichment of datasets by adding meta-data
- facilitating the creation of new databases by providing meta-information on concept lists
- enhancing the transparency of our research by providing a stable reference for all those who use concept lists in their research

The Concepticon

The Concepticon is an **attempt to link the many different concept lists ("Swadesh Lists")** which are used in the linguistic literature. In practice, **all entries from the various concept lists are linked to a *concept set*** as an intermediate way to reference the concepts. The Concepticon

- links **9611 concepts**
- from **51 concept lists**
- to **2206 concept sets** and
- defines **243 relations between the concept sets**.

List, Cysouw & Forkel (2015): Concepticon. Version 0.1, `http://concepticon.clld.org`.

The Concepticon: Concept Lists

A concept list is a **collection of concepts that is deemed interesting by scholars.** Minimally, it consists of an identifier for each concept which the lists contains, and a gloss by which the concept is referenced. The creator of a concept list is called a compiler. Each concept list is tight to one or more sources, it is given in one or more source languages and was compiled for one or more target languages. A description gives further information on each concept list in free, exclusively human-readable form.

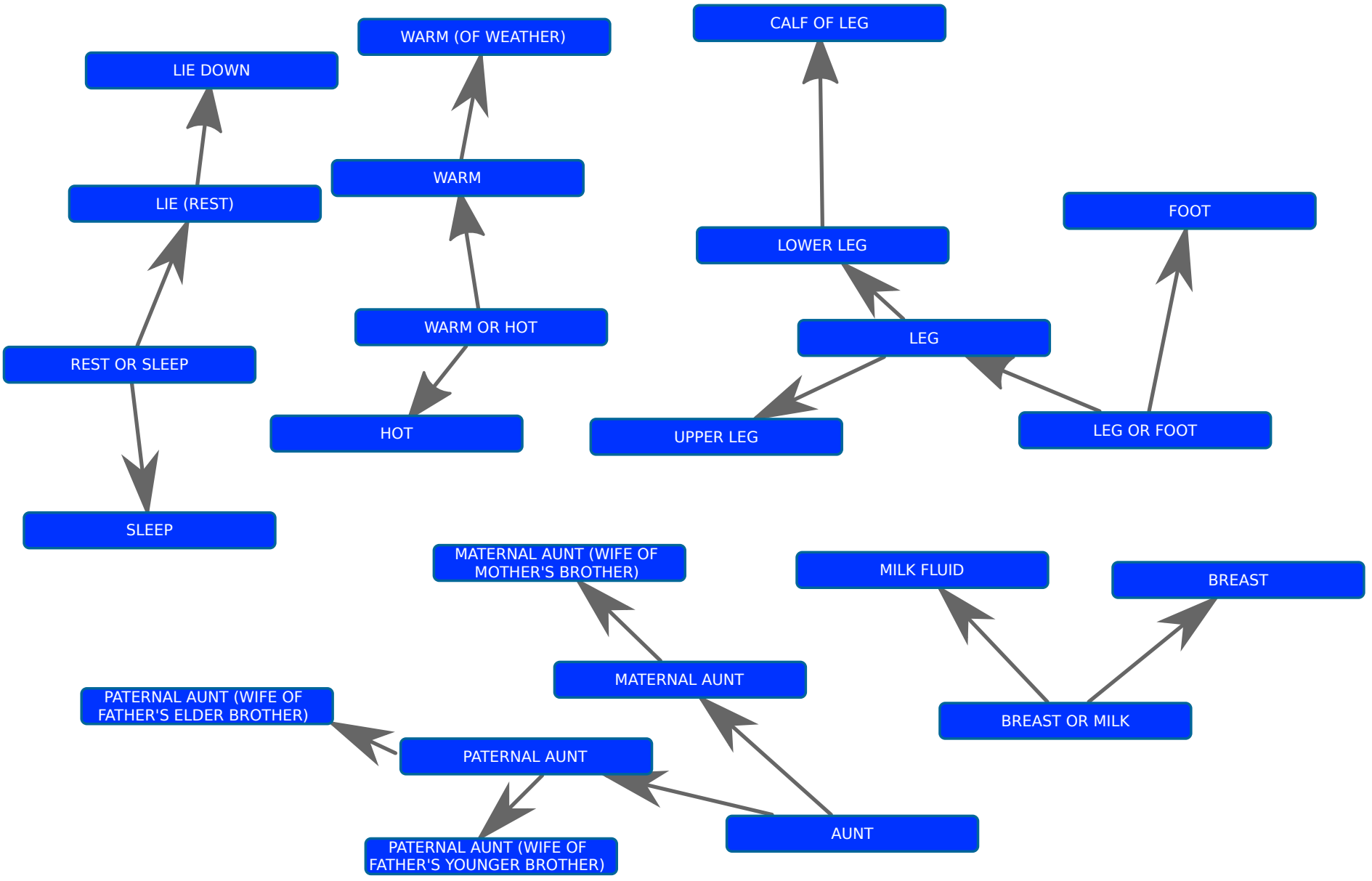
The Concepticon: Concept Sets

A concept set is a **collection of similar (ideally identical) concepts across the same or multiple concept lists.** Each concept set is represented in form of a gloss and in form of a definition and is defined by a unique numerical identifier. concept sets are further assigned to specific semantic fields (following closely those fields used in the WOLD project by Haspelmath & Tadmor 2009, `http://wold.clld.org`) and given an ontological category to help to order and identify the different concepts.

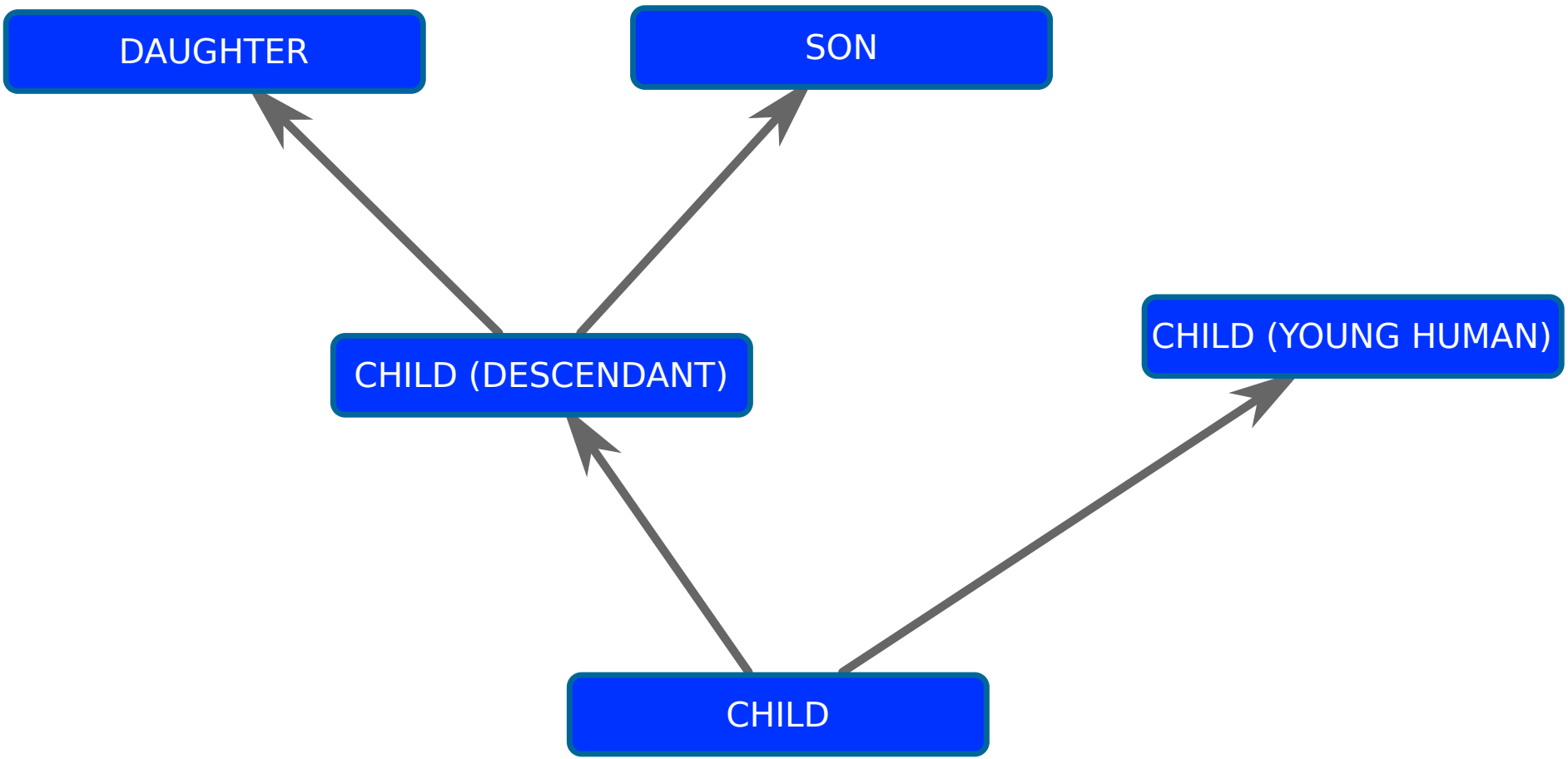
Concepticon: Concept Relations

To facilitate our workflow and to guarantee the comparability of concept lists even if they do not share concepts which are directly linked via our concept sets, **we define additional and very simple concept relations between concept sets** (*broader*, *narrower*, *similar*). Even if the concepts in two or more concept lists are not assigned to the same concept set, they can still be assigned to concept sets via concept relations.

Concepticon: Concept Relations



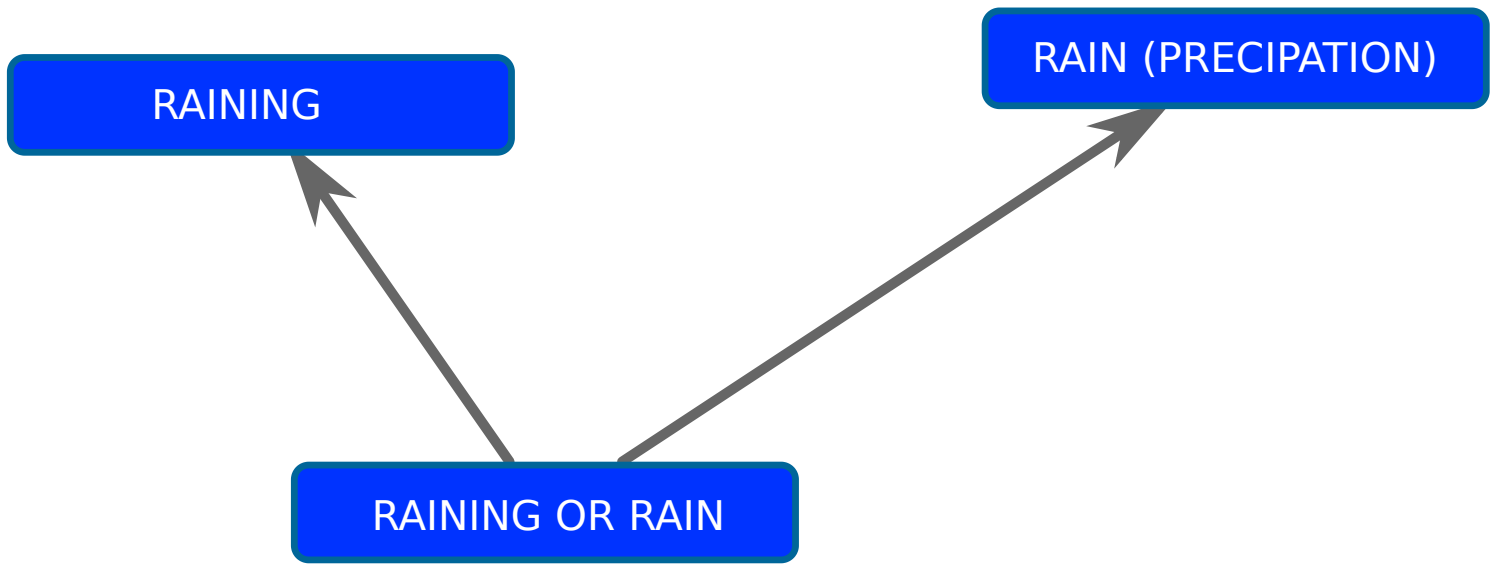
Examples: CHILD, RAIN, and BURN in “Swadesh Lists”



Examples: CHILD, RAIN, and BURN in “Swadesh Lists”

Compiler	Date	Items	CONCEPT	Concepticon
Blust	2008	210	child	CHILD
Chen	1996	200	孩子 / child	CHILD
Dunn	2012	207	child	CHILD
Leibniz	1768	128	infans	CHILD (YOUNG HUMAN)
Matisoff	1978	200	child/son	CHILD (DESCENDANT)
Swadesh	1950	215	child (son or daughter)	CHILD (DESCENDANT)
Swadesh	1952	200	child (young person rather than as relationship term)	CHILD (YOUNG HUMAN)
Tadmor	2009	100	child (kin term)	CHILD (DESCENDANT)
Wiktionary	2003	207	child (a youth)	CHILD (YOUNG HUMAN)

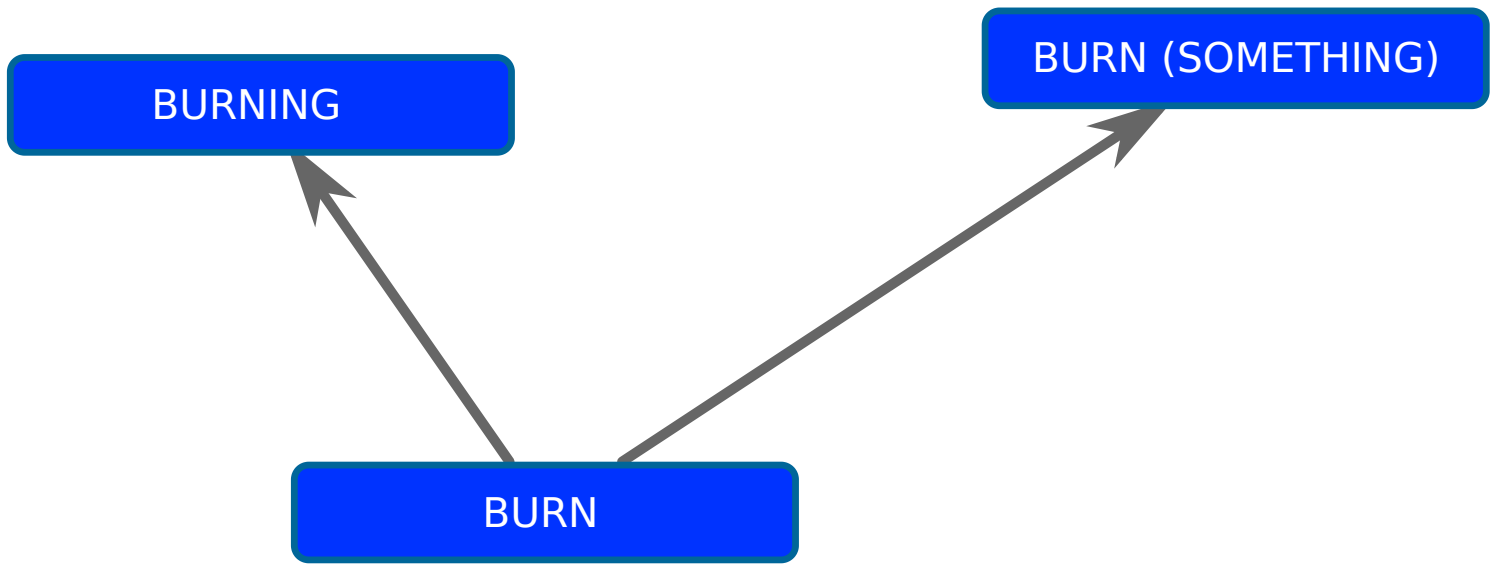
Examples: CHILD, RAIN, and BURN in “Swadesh Lists”



Examples: CHILD, RAIN, and BURN in “Swadesh Lists”

Compiler	Date	Items	CONCEPT	Concepticon
Blust	2008	210	rain	RAIN (PRECIPATION)
Chen	1996	200	雨 / rain	RAIN (PRECIPATION)
Dunn	2012	207	rain	RAINING OR RAIN
Leibniz	1768	128	pluvia	RAIN (PRECIPATION)
Matisoff	1978	200	rain	RAIN (PRECIPATION)
Swadesh	1950	215	rain	RAINING OR RAIN
Swadesh	1952	200	to rain	RAINING
Tadmor	2009	100	rain	RAIN (PRECIPATION)
Wiktionary	2003	207	to rain	RAINING

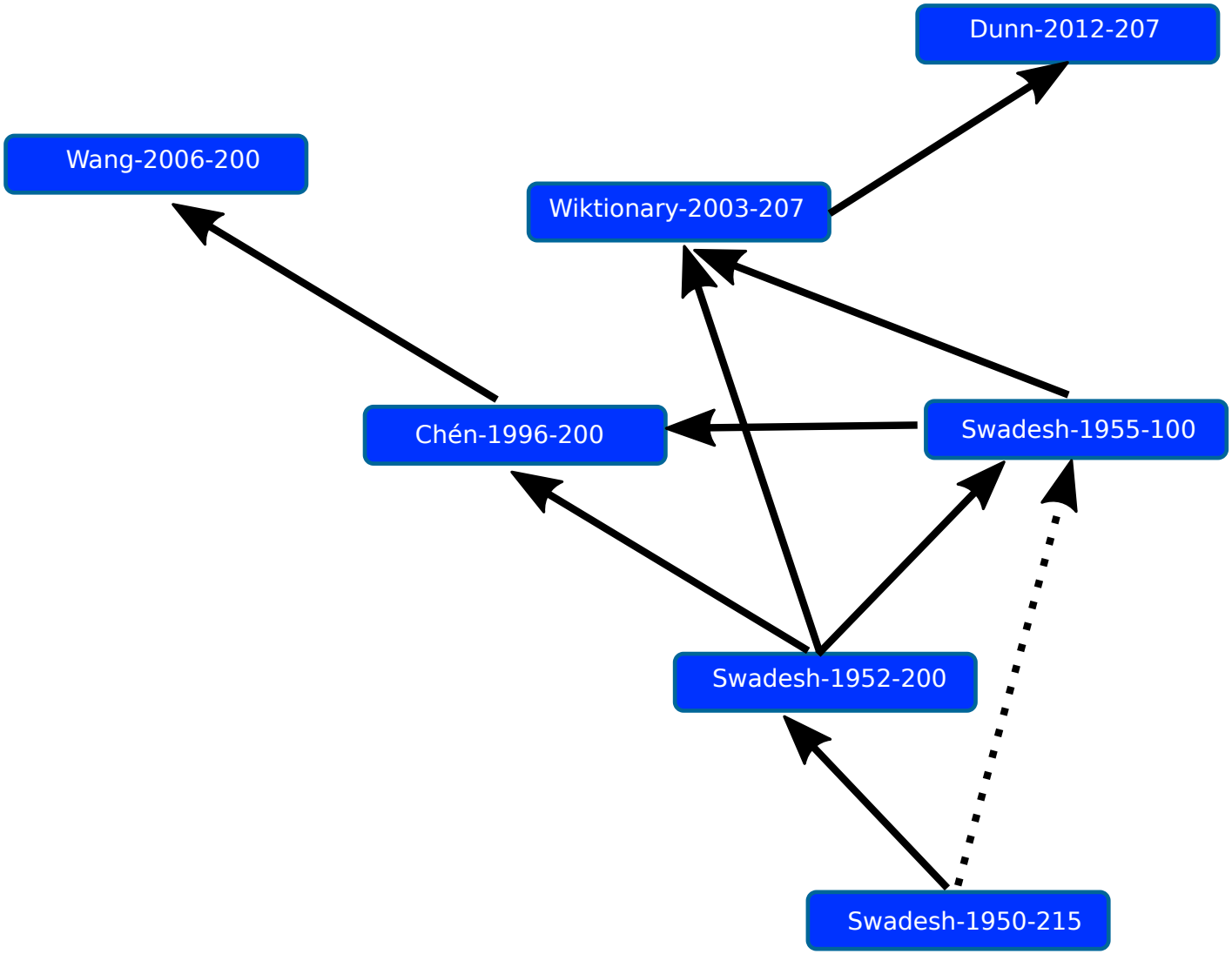
Examples: CHILD, RAIN, and BURN in “Swadesh Lists”



Examples: CHILD, RAIN, and BURN in “Swadesh Lists”

Compiler	Date	Items	CONCEPT	Concepticon
Blust	2008	210	to burn	BURN
Chen	996	200	烧 / burn	BURN
Dunn	2012	207	burn	BURN
Matisoff	1978	200	burn	BURN
Swadesh	1950	215	burn	BURN
Swadesh	1952	200	burn (intrans)	BURNING
Swadesh	1955	100	burn tr.	BURN (SOMETHING)
Tadmor	2009	100	to burn (intr.)	BURNING
Wiktionary	2003	207	to burn (intransitive)	BURNING

Examples: DULL, BLUNT, and STUPID



Examples: DULL, BLUNT, and STUPID

Compiler	Date	Items	CONCEPT	Concepticon
Blust	2008	210	dull, blunt	DULL
Chen	1996	200	呆，笨 / dull	STUPID
Dunn	2012	207	dull	DULL
Wang	2006	200	笨（ 不聪明 ）/ dull	STUPID
Swadesh	1952	200	dull (knife)	DULL
Wiktionary	2003	207	dull (as a knife)	DULL

Directions

Increasing the Data Basis

- mapping further concept lists
- inviting scholars to contribute

Refining the Data

- glosses and definitions
- concept relations
- meta-data (more links, translation of glosses)

Refining the Workflow

- refine scripts for automatic mapping
- formalize workflow for manual mapping
- decide open questions (see next slides)

Discussions

Current Workflow

1. Digitization of a concept list: OCR or type of or copy-paste a concept list from the literature into a TSV-file.
2. Preparation of the concept list: Translate glosses into English if they are only given in another language and search for useful ways to link the concept lists (URLs, for example) and add them to the TSV-file as separate columns.
3. Mapping of the concept list to the Concepticon: Start by using an automatic method for fuzzy mapping and then refine the automatic mapping manually.
4. Updating the Concepticon application. In case of mapping difficulties:
   - Add a new concept set to the Concepticon (along with gloss and definition), if a concept cannot be mapped to any concept set.
   - Define, if needed, concept relations between the new concept set and the existing ones.

Alternative Proposal: Background

- The Concepticon cannot offer a complete semantic analysis.
- We only can provide an approximate matching between concepts.

Alternative Proposal: Proposal

- Concept sets should be semantically disjoint.
- So: there should never be semantic overlap between concept sets.
- The (editorial) decision about the boundaries between sets is not trivial.

Alternative Proposal: Consequences

- Concept sets will often be semantically somewhat diverse (e.g. “MARRY”).
- Some Concepts will only be a subset of a concept set (e.g. “marry a woman” vs. “marry a man”).
- Some Concepts will be linked to multiple concept sets (e.g. “hand/arm” to “HAND” and “ARM”).
- Sometimes multiple concepts from the same concept list will be linked to the same concept set.

Alternative Proposal: Editorial Work

- The definitions of the concept sets need to be checked for overlap.
- When overlap exists,
  - either the sets have to be merged to an overarching concept set,
  - or the definitions have to be changed to be disjoint.
- In the future it is possible that new insights suggest the splitting of a concept set.
  - then all links will have to be reconsidered.

Summary

Aspect	Current Proposal	Alternative Proposal
concept sets	allow for overlap	keep them disjoint
concept relations	needed to guarantee comparability	can be ignored
links	assign each concept to one concept set	allow to assign one concept to multiple concept sets
compatibility	can be automatically converted to the other	cannot be automatically converted
mapping	no re-editing required, but constant editing of concept sets and relations	re-editing of all concept lists constantly required, adding of concept sets restricted

Thanks to

- Martin Haspelmath for helpful discussions and practical and “ideological” support
- our student assistants: Viola Kirchhoff, Frederike Urke, and Sebastian Nicolai

Thanks to

- Martin Haspelmath for helpful discussions and practical and “ideological” support
- our student assistants: Viola Kirchhoff, Frederike Urke, and Sebastian Nicolai

Thank You for Listening!

Thanks to

- Martin Haspelmath for helpful discussions and practical and “ideological” support
- our student assistants: Viola Kirchhoff, Frederike Urke, and Sebastian Nicolai

Discussion is Open!