Graph-based approaches to meaning and language comparison

Volker Gast

Friedrich Schiller University Jena
Palacký University Olomouc

5 Feb 2018

Why graphs . . . ?

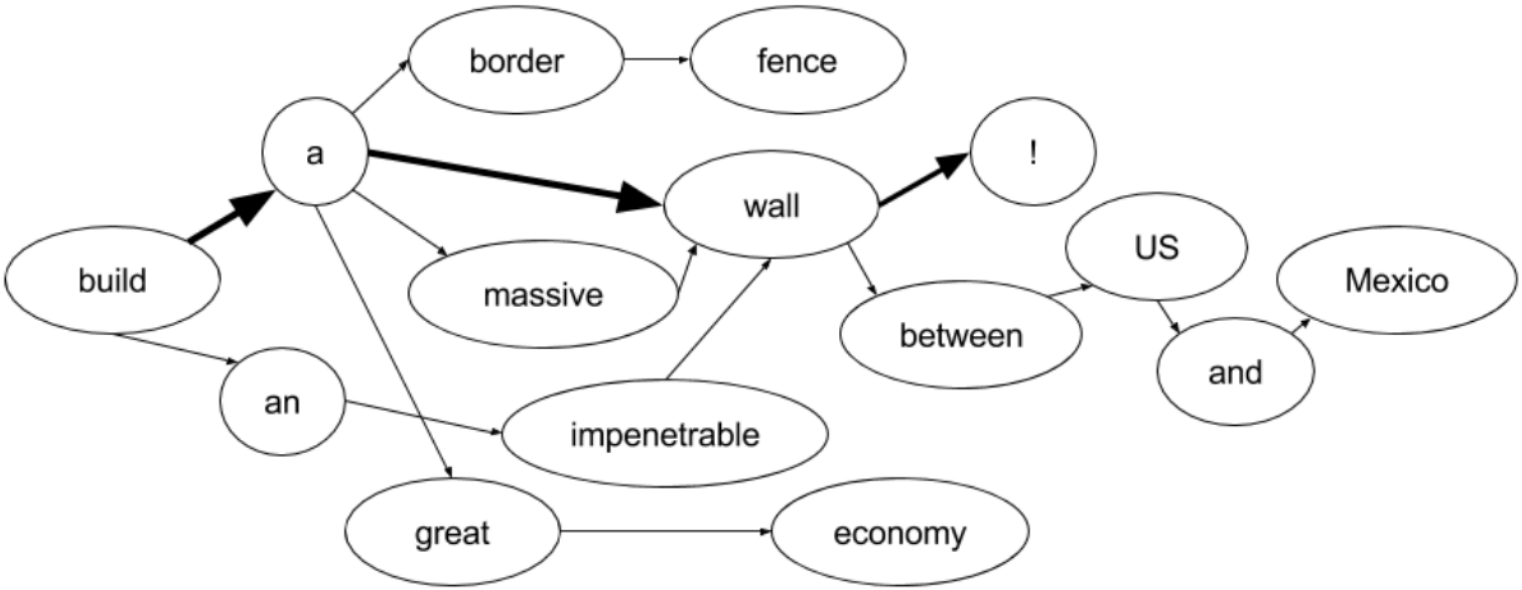
- Linguistic systems of all types, as well as 'linguistic behaviour' more generally speaking, can be described in terms of sets of elements (e.g. words) and relations holding between these elements.
- Graphs are just that – sets of elements (nodes, vertices) and edges (pairs of nodes/vertices, i.e. relations).

  $G = <\mathrm{N}, \mathrm{E}>$,

  $N = \{n_1, n_2 \ldots n_n\}$,

  $E = \{e_1, e_2 \ldots e_n\}$, $\mathrm{e}_i = <\mathrm{n}_j, \mathrm{n}_k>$
- Examples: Syntactic trees (directed acyclic graphs), lexical networks (WordNet), Markov models, etc.
- Linguistic systems as 'networks' $\rightarrow$ psychologically plausible, computational implementation is straightforward.
- Both the concept of a 'graph' and the computational ressources can be very useful for a wide variety of linguistic tasks.

Automatic Donald Trump (Markov model)



https://filiph.github.io/markov/

Overview

Types of graphs discussed in this talk

- Distribution graphs
  - Impersonals: mapping the potential distribution of a linguistic element
  - Scalar additive operators: mapping the actual distribution of linguistic elements
- Translation graphs
  - Adverbials of immediate posteriority: mapping translation behaviour
- Information graphs
  - Mapping the information conveyed in communicative events, and the processing of this information
  - Mapping the temporal structure of texts

Distribution graphs, aka ('traditional') 'semantic maps'

- 'Traditional' semantic maps as a way to represent patterns of polysemy and carry out crosslinguistic comparison.
- Prominent (early) examples: indefinites[1], modals;[2] see Georgakopoulos and Polis (forthcoming)[3] for a recent survey.

[1] M. Haspelmath (1997). *Indefinites*. Oxford: Oxford University Press.

[2] J. van der Auwera and V. Plungian (1998). "Modality's semantic map". In: *Linguistic Typology* 2, pp. 79–124.

[3] T. Georgakopoulos and S. Polis (forthcoming). "The semantic map model. State of the art and future avenues for linguistics research". In: *Language and Linguistics Compass*.

Indefinites

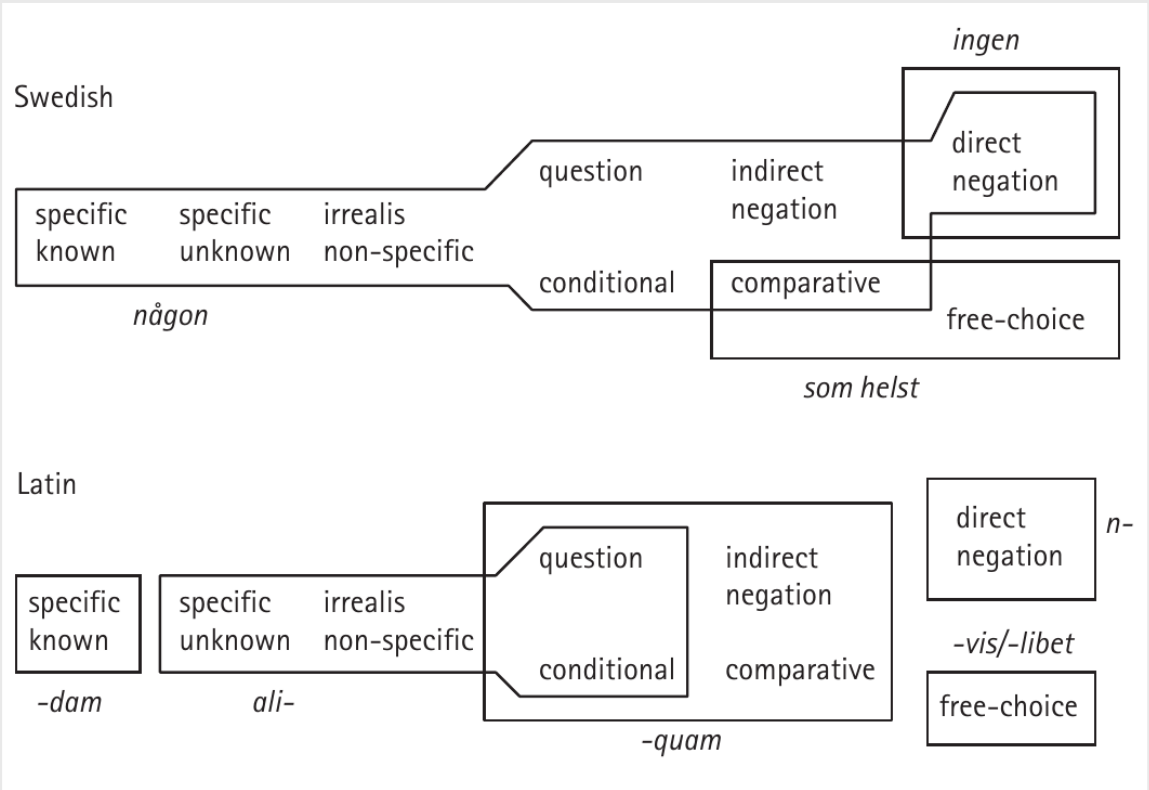


Figure 9.6. Indefinite pronouns in Swedish and Latin (Haspelmath 1997: 68–9)

Distribution graphs and distance-based maps

- Distance-based maps: Similarity (mostly in terms of distribution) is represented as proximity in two-dimensional space.
- Are distance-based maps the better way of representing (crosslinguistic) distributions?

> "...the semantic map model, while theoretically well-motivated in typology, is not mathematically well-defined or computationally tractable, making it impossible to use with large and highly variable datasets."

(Croft and Poole, 2008, p. 1)[4]

[4] W. Croft and K. Poole (2008). "Inferring universals from grammatical variation: Multidimensional scaling for typological analysis". In: *Theoretical Linguistics* 34.1, pp. 1–37.

My point of view

- Distribution graphs and 'distance-based semantic maps' are different types of things.
- Distance-based semantic maps are visualizations of multi-dimensional data (implying simplification, loss of information). They can be used as bottom-up ways to explore multi-dimensional data.
- Distribution graphs are hypotheses about restrictions on linguistic systems.
- In distribution graphs, the meaning is (ideally) given and distribution is shown, whereas in distance-based semantic maps distribution is given and meaning is inferred.
- Both methods are valuable and can complement each other (and both methods are mathematically well-defined).

Impersonals and distribution graphs

- Impersonal pronouns: Engl. *one*, Fr. *on*, Germ. *man* etc.

(1) **One** only lives once.

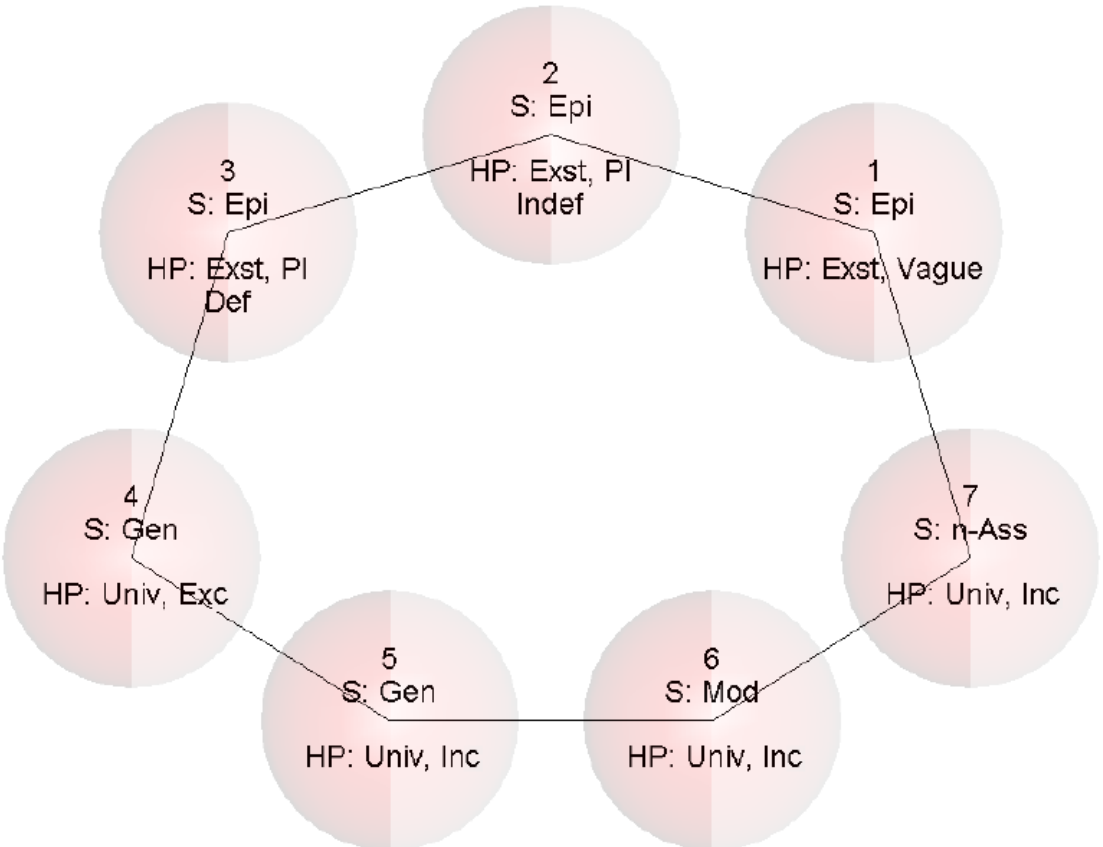
(2) **On** ne vit qu'une fois.

(3) **Man** lebt nur einmal.

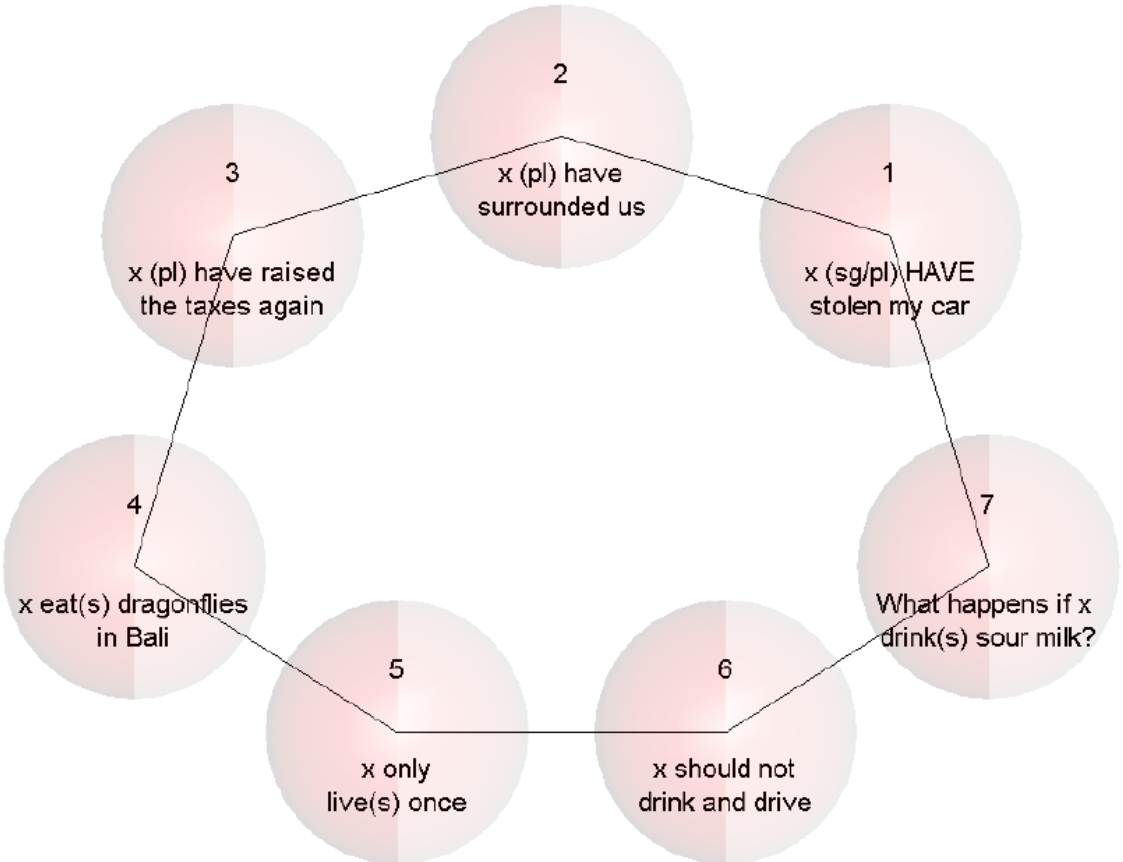
- Each node in a graph corresponds to a specific meaning-context combination, defined on the basis of theoretical work.
- Meanings and context types are represented by features, e.g.:[5]
  - episodic vs. generic event descriptions
  - veridicality (veridical, non-veridical)
  - type of quantification (existential, universal)
  - inclusion / exclusion of addressee

[5] V. Gast and J. van der Auwera (2013). "'Towards a distributional typology of human impersonal pronouns, based on data from European languages". In: *Languages Across Boundaries. Studies in Memory of Anna Siewierska*. Ed. by D. Bakker and M. Haspelmath. Berlin: de Gruyter Mouton, pp. 119–158.

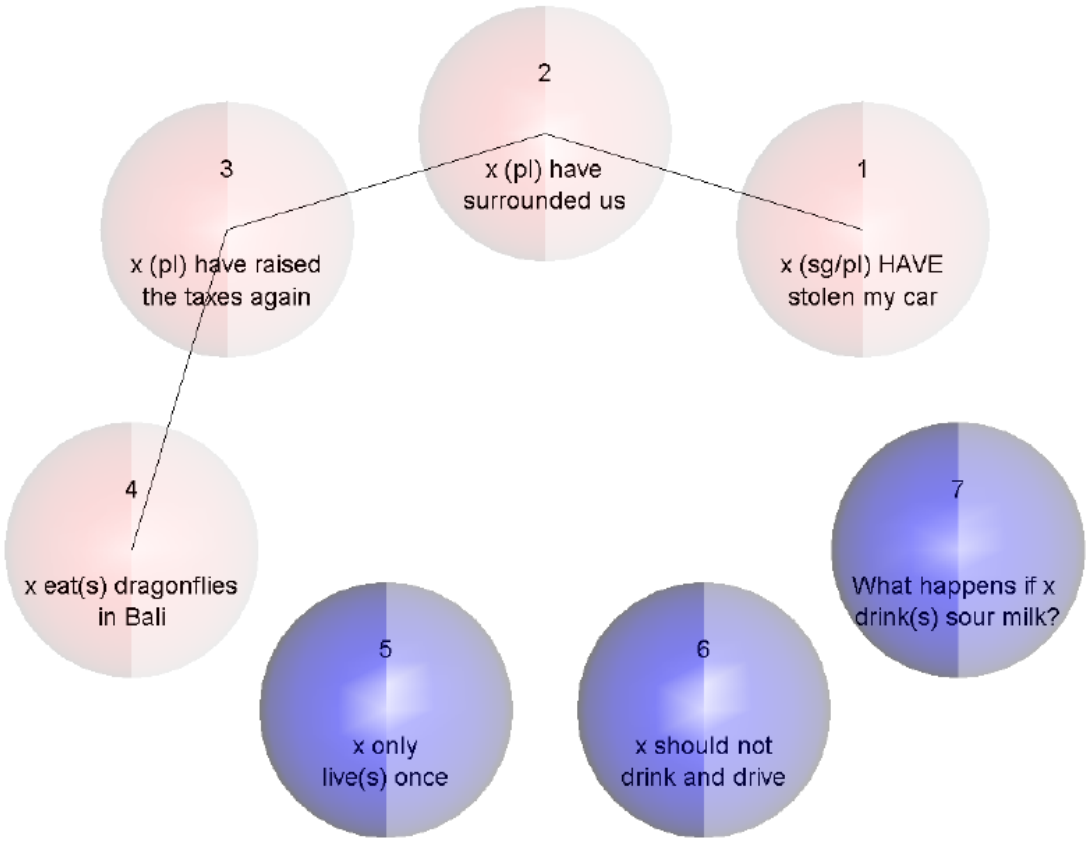
Impersonals: Features



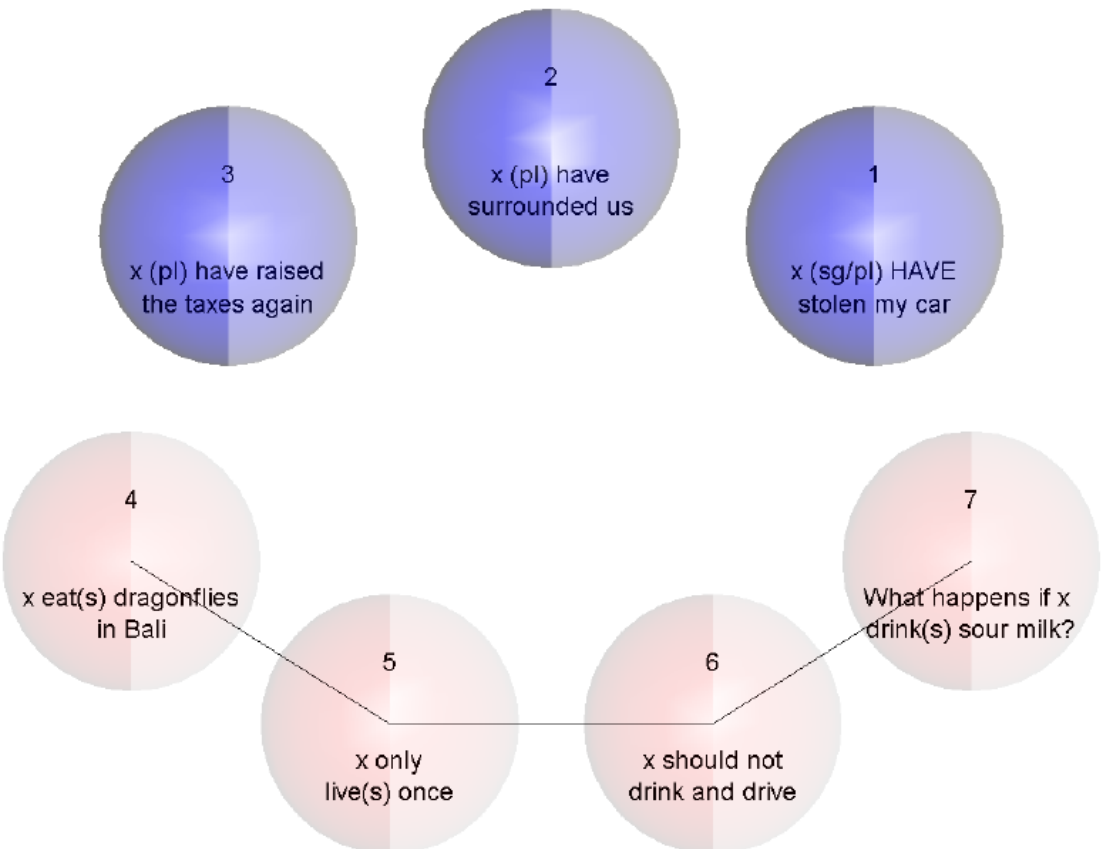
Impersonals: Diagnostic contexts



Impersonals: Engl. *they*



Impersonals: Ital. *si*



Impersonals: Examples of Ital. *si*

(4) In Spagna si cena tardi.
'In Spain they eat late.'

(5) Si vive solo una volta.
'One/you only live(s) once.'

(6) Non si deve bere alla guida.
'One/you shouldn't drink and drive.'

(7) Cosa succede se si beve del latte scaduto?
'What happens if you/one drink(s) sour milk?'

Scalar additive operators and distribution graphs

- Scalar additive operators: Engl. *even*, Fr. *même*, Germ. *sogar* etc.
- Case study: use of three scalar additive particles in German as translation equivalents of *even*.[6]
- Engl. *even* is translated in various ways, mostly with *sogar*, *selbst* or *auch*.
- What translation is found in what context types / under what conditions?

[6] V. Gast (2017). "The scalar operator *even* and its German equivalents: Pragmatic factors determining the use of *auch*, *selbst* and *sogar*". In: *Focus on Additivity. Multifaceted views on Focusing Modifiers*. Ed. by A.-M. De Cesare. Amsterdam: Benjamins, pp. 201–234.

Some examples

(8) **Even** with a [$_F$ moderate] level of reduction, effects on communities are severe.

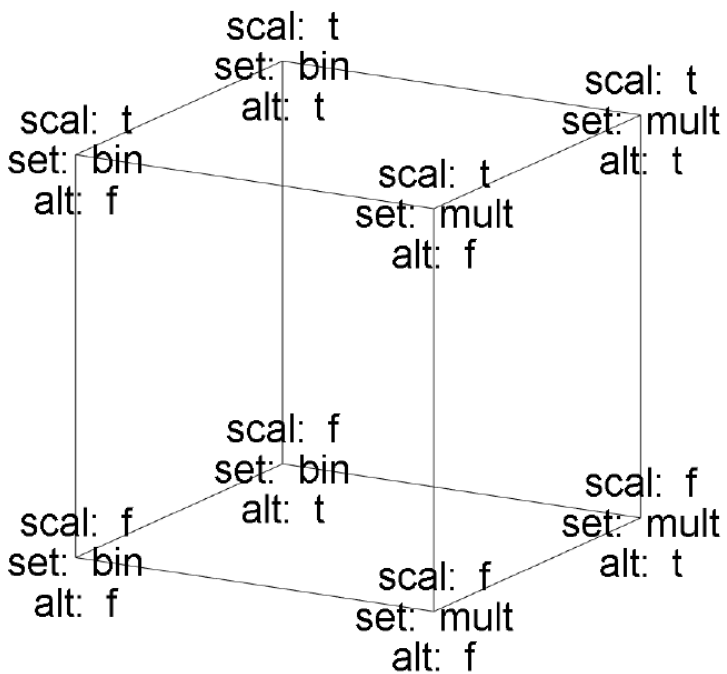
(9) The key result has been the adoption of a joint declaration on agricultural expenditure, allowing the Commission to submit a letter of amendment to its preliminary draft budget **even** [$_F$ outside] the procedural provisions of the financial regulation.

(10) A remarkable part of this report is dedicated to the countries in Eastern Europe that might in the future become members of the EU family, namely Ukraine, Moldova and **even** [$_F$ Belarus].

Scalar additive operators: Parameters of variation

- How many focus alternatives are there (two/bin, more than two/mult)?
- What type of contrast is established (inherently/contextually scalar)?
- Are focus alternatives mentioned in the relevant sentence?

A three-dimensional feature grid (2×2×2)



The distribution of *selbst*

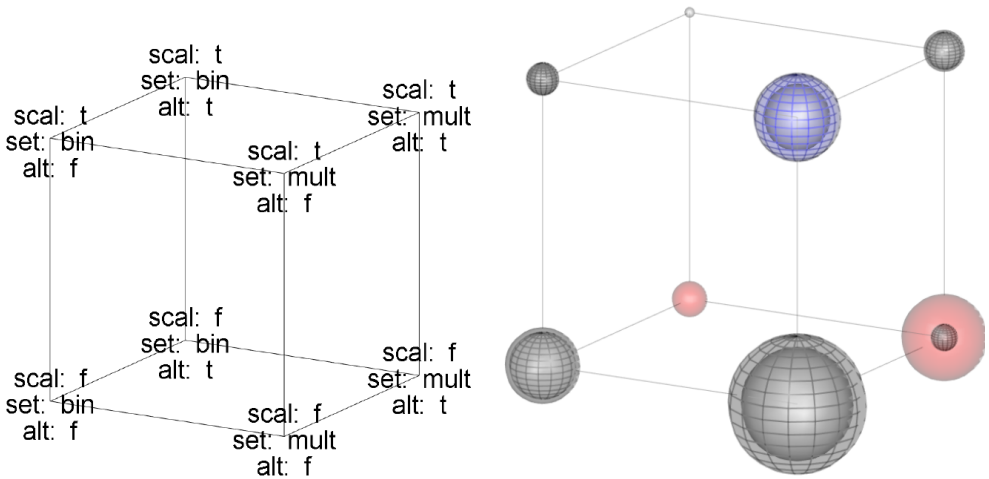


Figure 11: FCF-grid for *selbst*

(11) Voluntary modulation of 20%, top-slicing the single farm payment, will mean that our farmers are likely to be 20% worse off than **even** [$_F$ their Welsh, Scottish and Ulster counterparts], let alone the French.

(12) Die freiwillige Modulation in Höhe von 20%, durch die Teile der einheitlichen Betriebsprämie zur Finanzierung anderer Zwecke gekürzt werden, bedeutet, dass unsere Landwirte im Vergleich **selbst** [$_F$ zu den Landwirten in Wales, Schottland und Ulster], von den französischen Landwirten ganz zu schweigen, 20% weniger verdienen werden.

The distribution of *sogar*

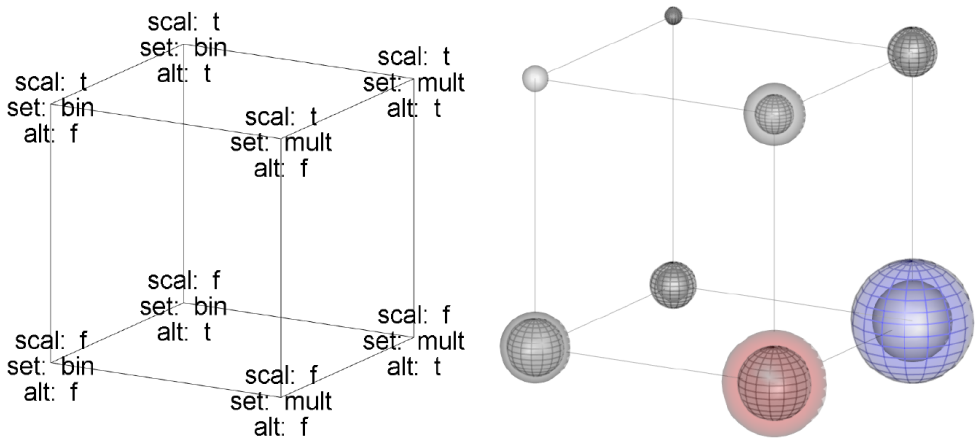


Figure 10: FCF-grid for *sogar*

(13) A remarkable part of this report is dedicated to the countries in Eastern Europe that might in the future become members of the EU family, namely Ukraine, Moldova and **even** [$_F$ Belarus].

(14) Ein erheblicher Teil dieses Berichts ist den Ländern Osteuropas gewidmet, die in Zukunft Mitglied der EU-Familie werden könnten, nämlich der Ukraine, Moldawien und **sogar** [$_F$ Belarus].

The distribution of *auch*

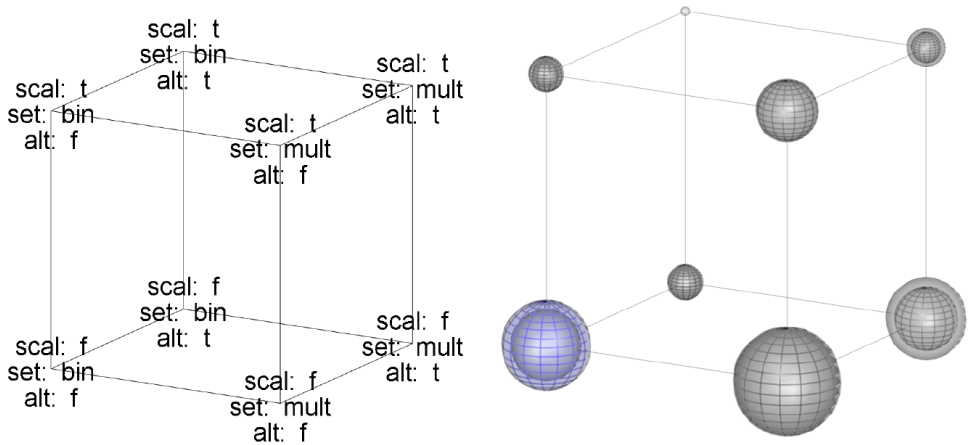


Figure 9: FCF-grid for *auch*

(15) The key result has been the adoption of a joint declaration on agricultural expenditure, allowing the Commission to submit a letter of amendment to its preliminary draft budget **even** [$_F$ outside] the procedural provisions of the financial regulation.

(16) Das wichtigste Ergebnis war die Annahme einer gemeinsamen Erklärung zu den Agrarausgaben, wodurch die Kommission dem Haushaltsvorentwurf **auch** [$_F$ außerhalb] der Verfahrensvorschriften für die Finanzordnung ein Berichtigungsschreiben hinzufügen konnte.

Translation Graphs: Adverbials of immediate posteriority

- Study of 'adverbials of immediate posteriority', e.g. Fr. *immédiatement*, *tout de suite*, Germ. *sofort*, *gleich*.

(17) a. Faites demi-tour **immédiatement**, s'il vous plaît!

b. Drehen Sie **sofort** um.

(18) a. Quand je t'ai vue avec tes fleurs, **tout de suite**, j'ai eu envie d'être heureux.

b. Als ich dich das erste Mal sah, wollte ich **sofort** glücklich mit dir werden.

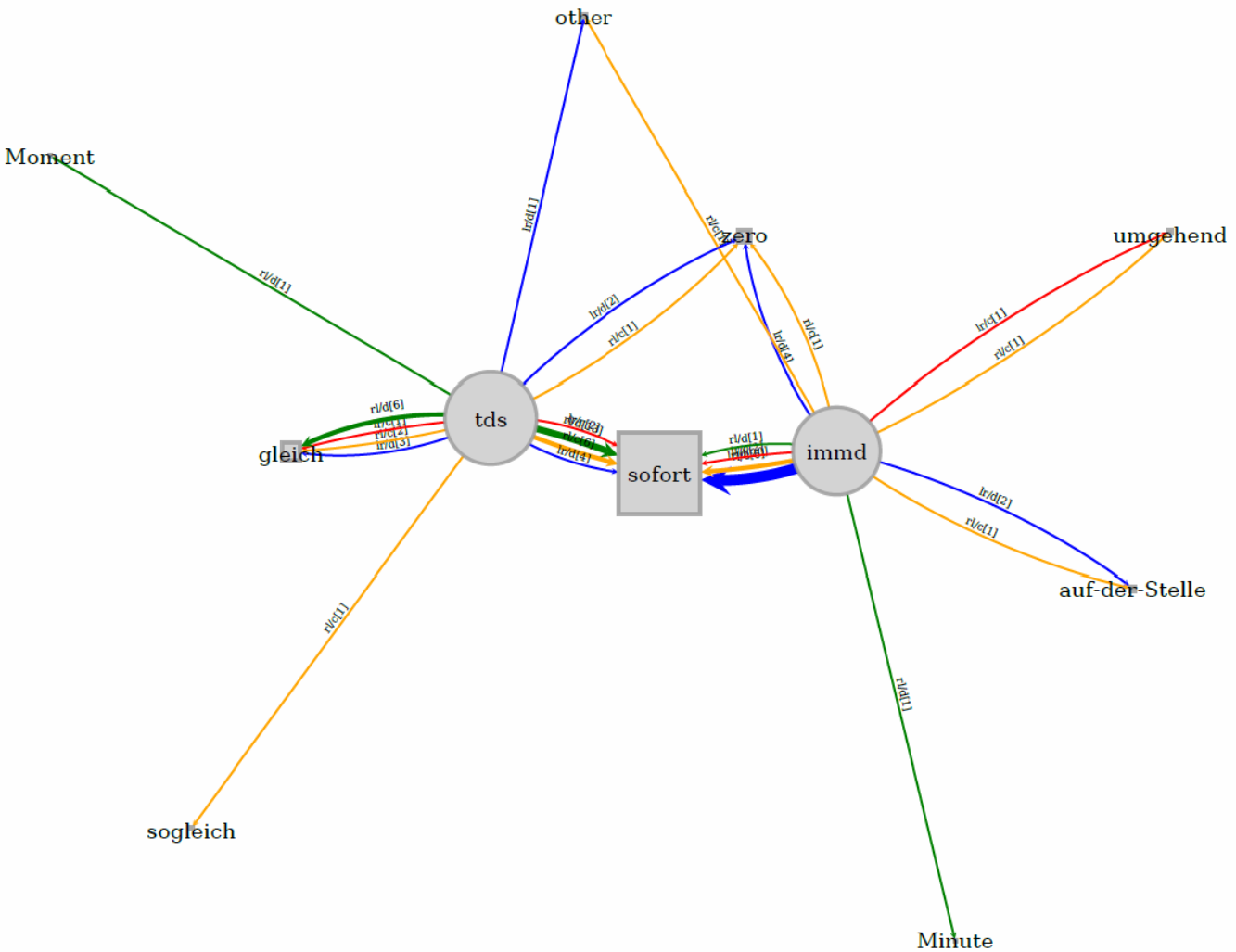
Parameters of variation

- Adverbials of immediate posteriority vary along several distributional parameters.
- Two parameters were shown to be particularly important in our studies, the reference point (deictic vs. chronological) and the 'direction of fit' (word-to-world, world-to-word).[7,8]
- Deictic adverbials refer to the moment uf utterance ('right now'), chronological ones to a preceding event ('right after *e*').
- Direction of fit:
  - word-to-world: representative speech acts etc.
  - world-to-world: declarations, directives

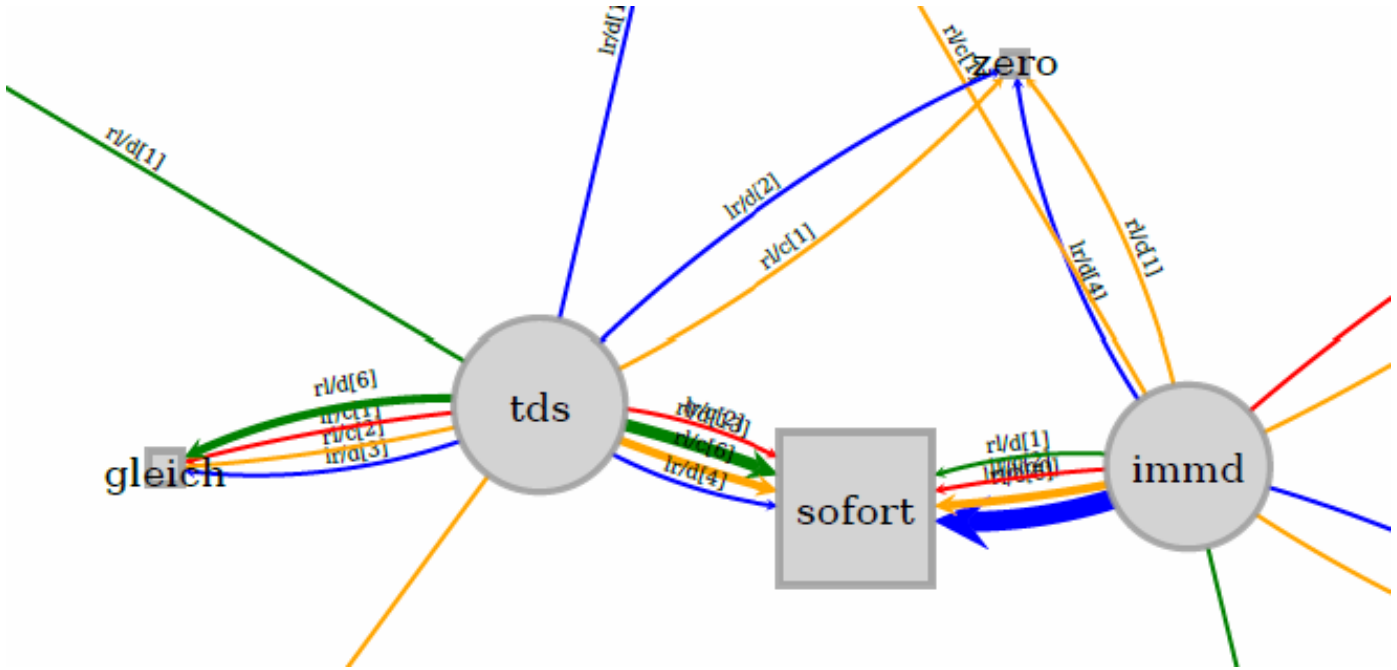
[7] V. Atayan et al. (forthcoming). "Ausdrucksformen der unmittelbaren Nachzeitigkeit in Originalen und Übersetzungen: Eine Pilotstudie zu den deutschen Adverbien *gleich* und *sofort* und ihren Äquivalenten im Französischen, Italienischen, Spanischen und Englischen". In: *Translation und Linguistik*. Ed. by B. Ahrens et al. Berlin: Frank & Timme.

[8] V. Gast et al. (forthcoming). "Unmittelbare Nachzeitigkeit im Deutschen und Französischen: Eine Studie auf Grundlage des OpenSubtitles-Korpus". In: *Comparatio delectat III. Akten der VIII. Internationalen Arbeitstagung zum romanisch-deutschen und innerromanischen Sprachvergleich*. Ed. by C. Konecny et al. Frankfurt: Lang.

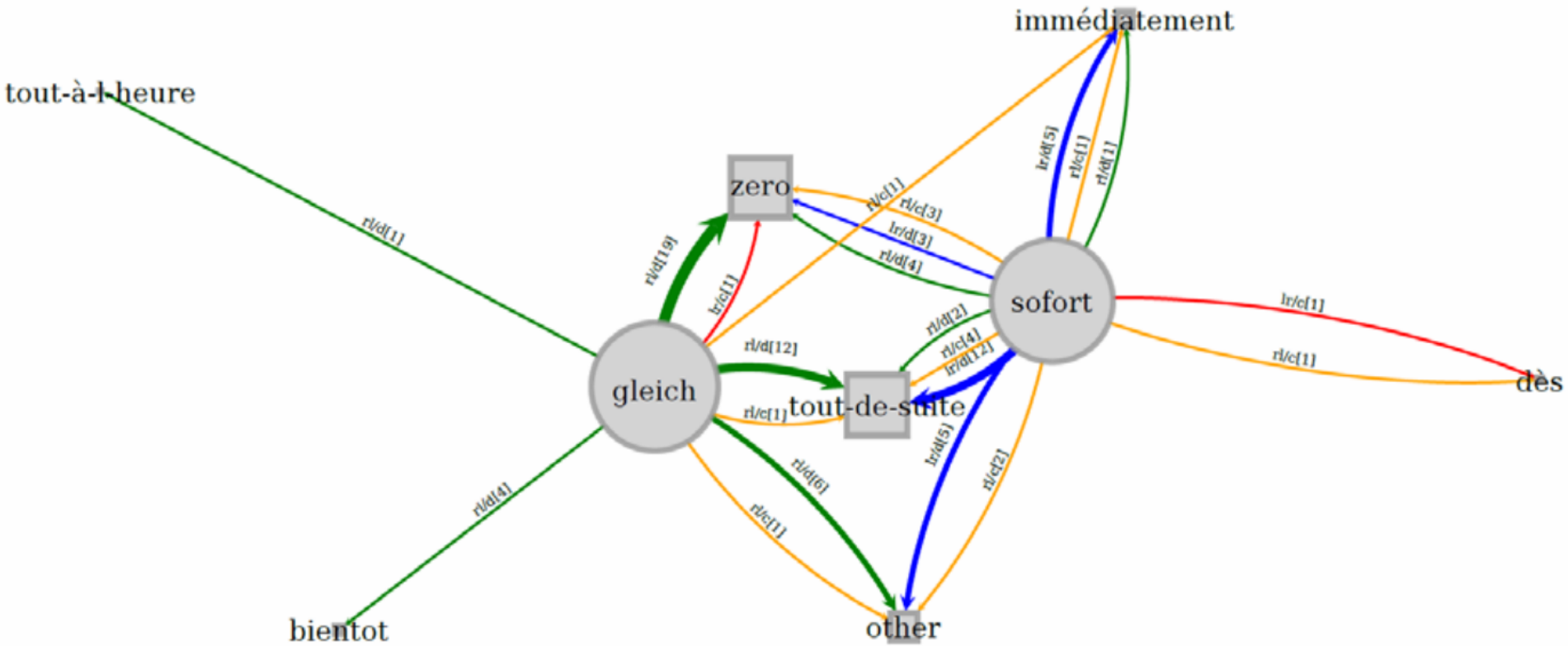
Representing translators' behaviour Fr. → Germ.



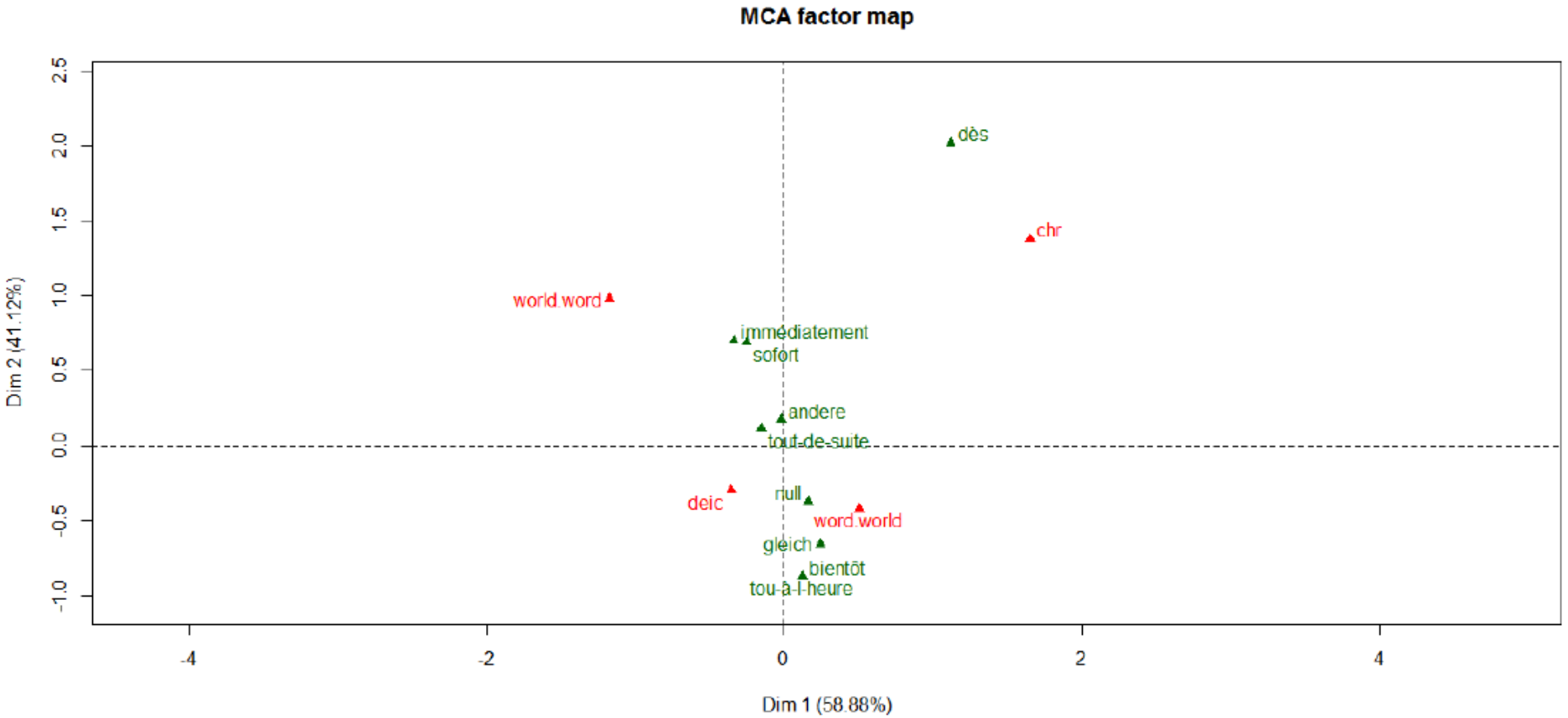
Representing translators' behaviour Fr. → Germ.



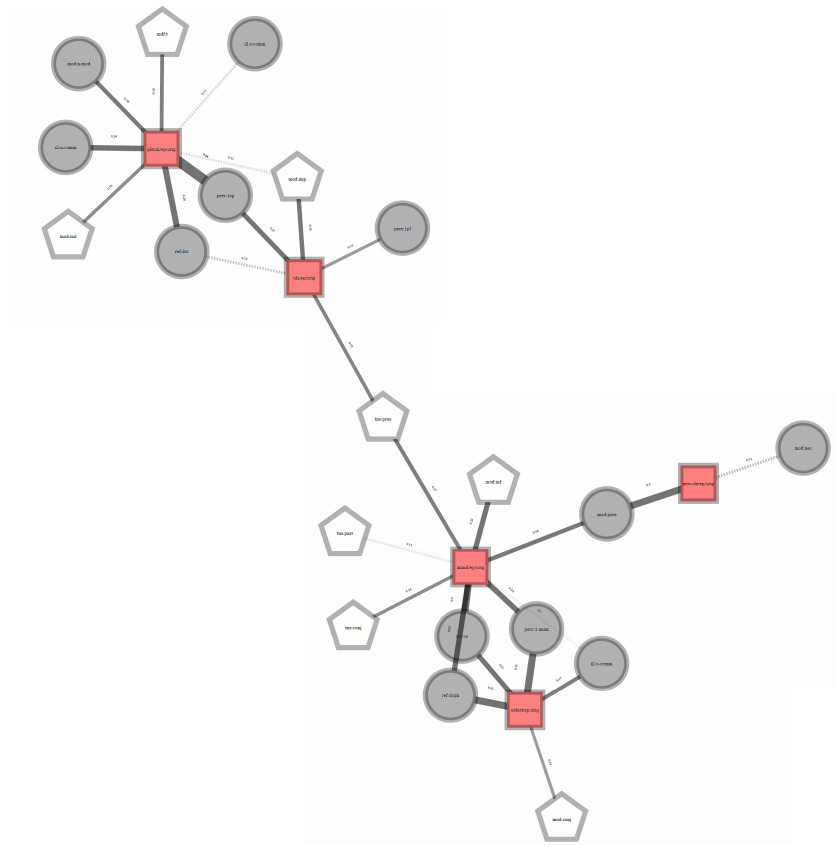
Representing translators' behaviour Germ. → Fr.



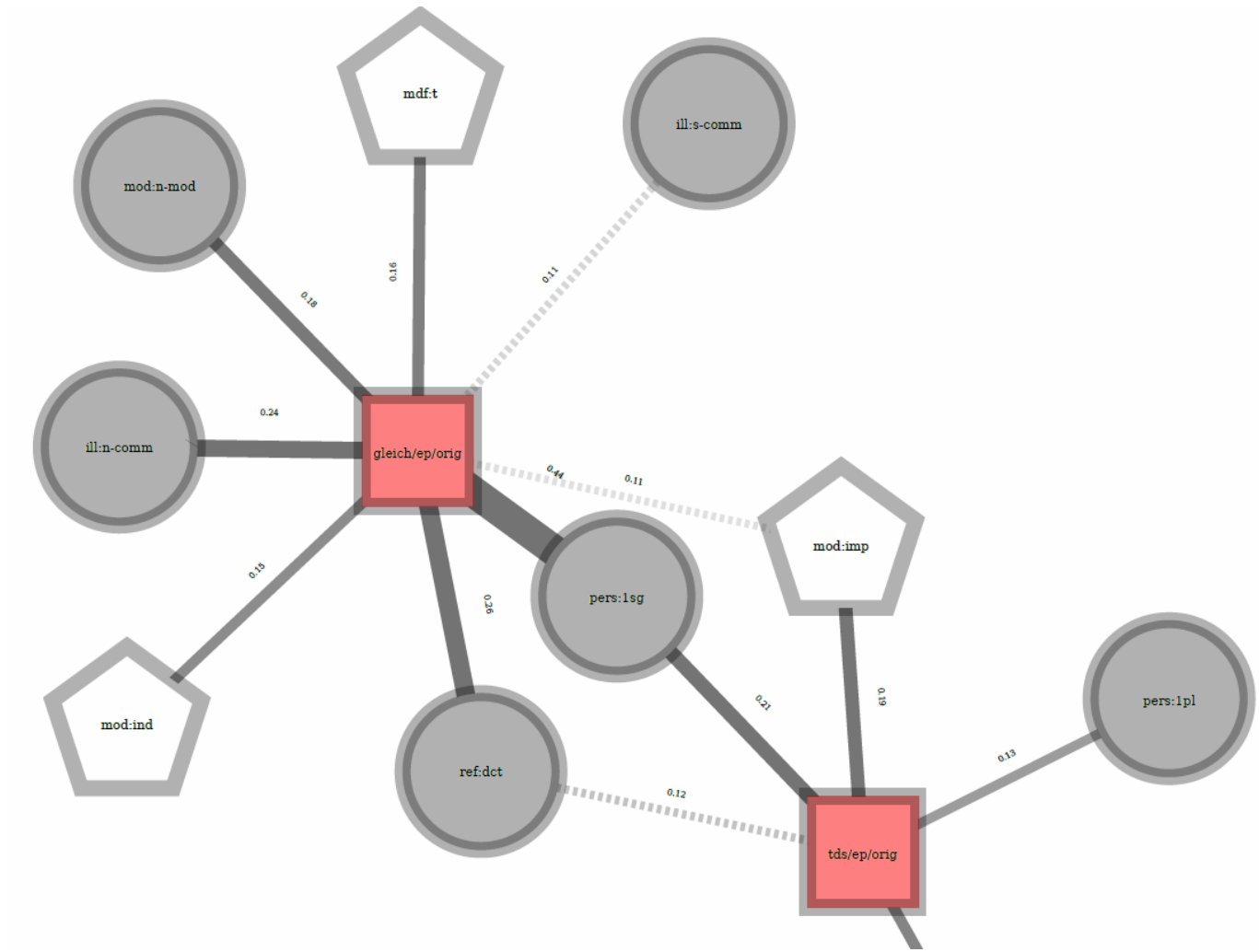
Adverbials of immediate posteriority: MCA



A more global picture



A more global picture (zoom)



Information processing

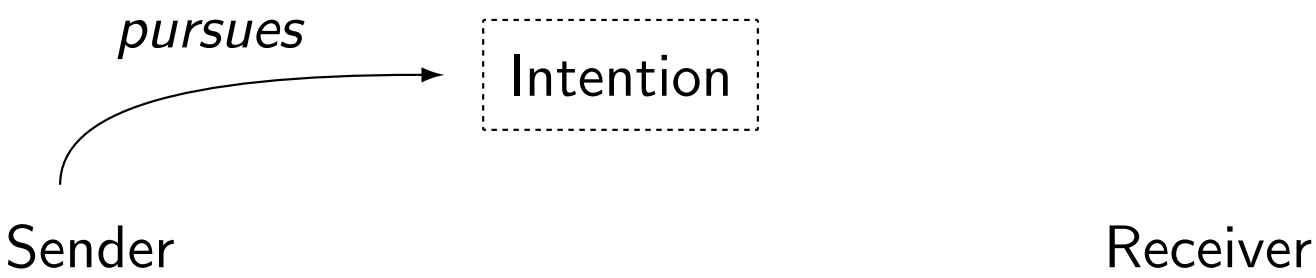
- Assumption: We have no direct access to 'reality' but can only talk about 'models', i.e. simplified representations of the world, formed on the basis of sensory perception → 'information models'.
- Each one of us has their own ('private') information model, and we create 'meta-models' of our conversation partners.
- Information models are graphs; the world is represented as a set of entities (with specific properties/belonging to specific categories) and relations holding between these entities.
- A communicative event provides an input to an information model, which is subsequently processed, modifying the model in accordance with the inferential processes triggered (cf. Relevance Theory).
- Utterances are defined as actions intended to update the addressee's information model.

An interactional model of communication

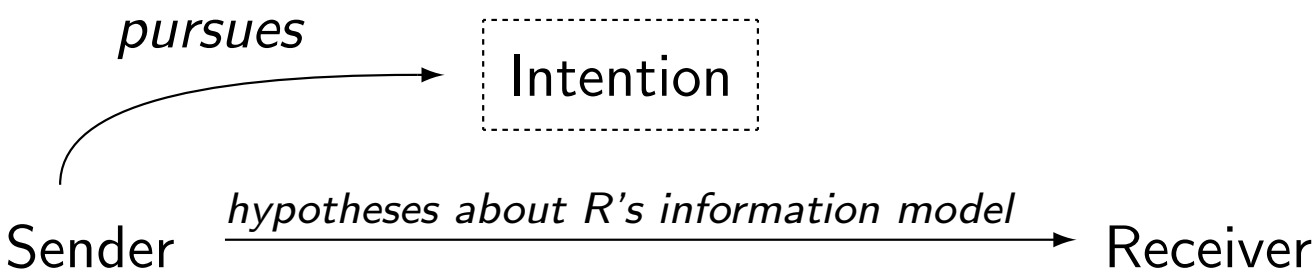
Sender

Receiver

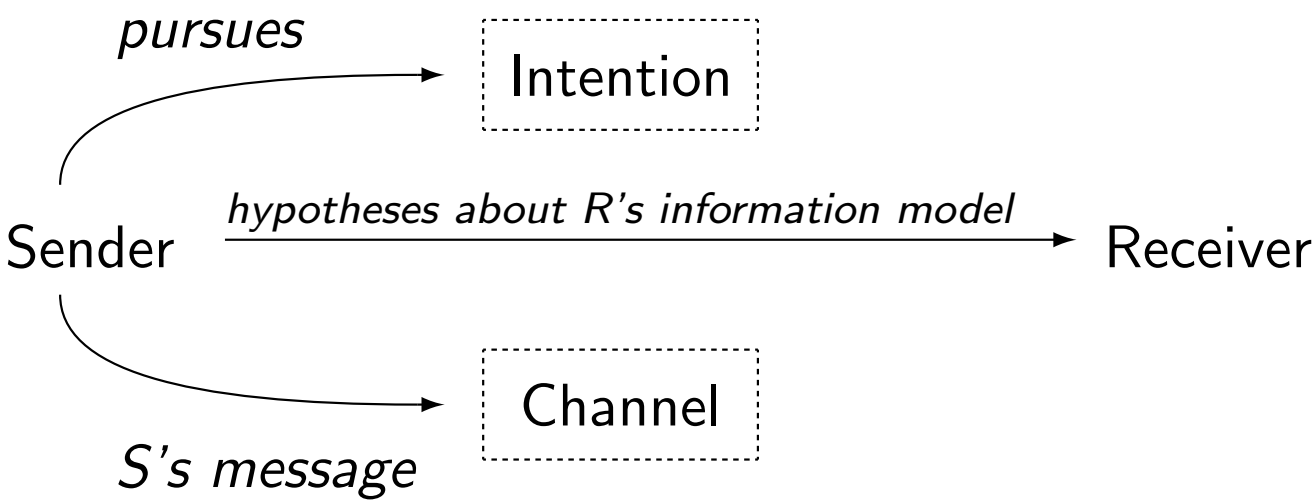
An interactional model of communication



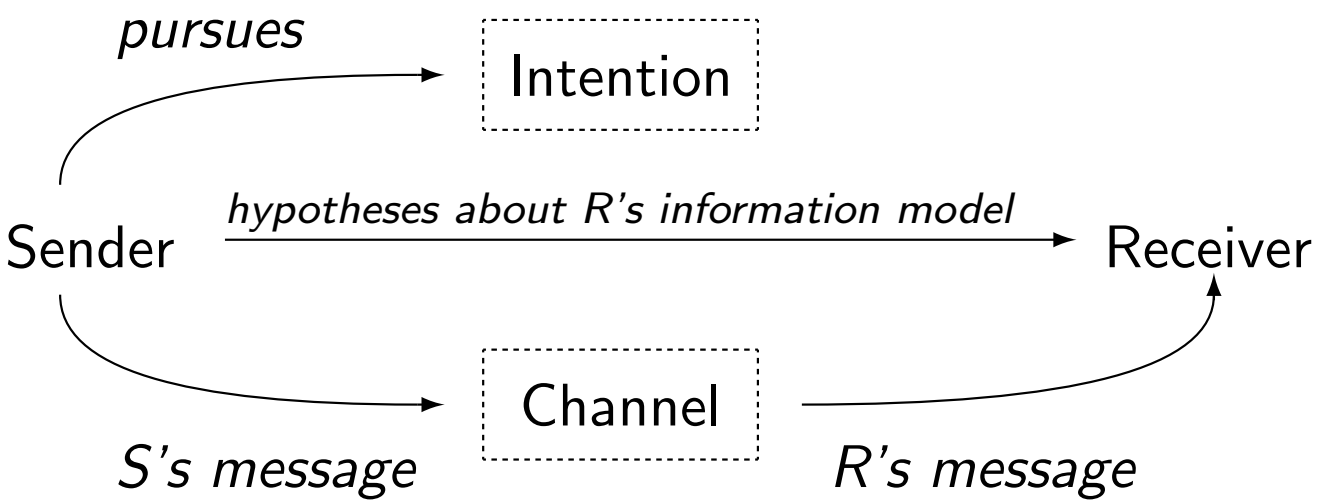
An interactional model of communication



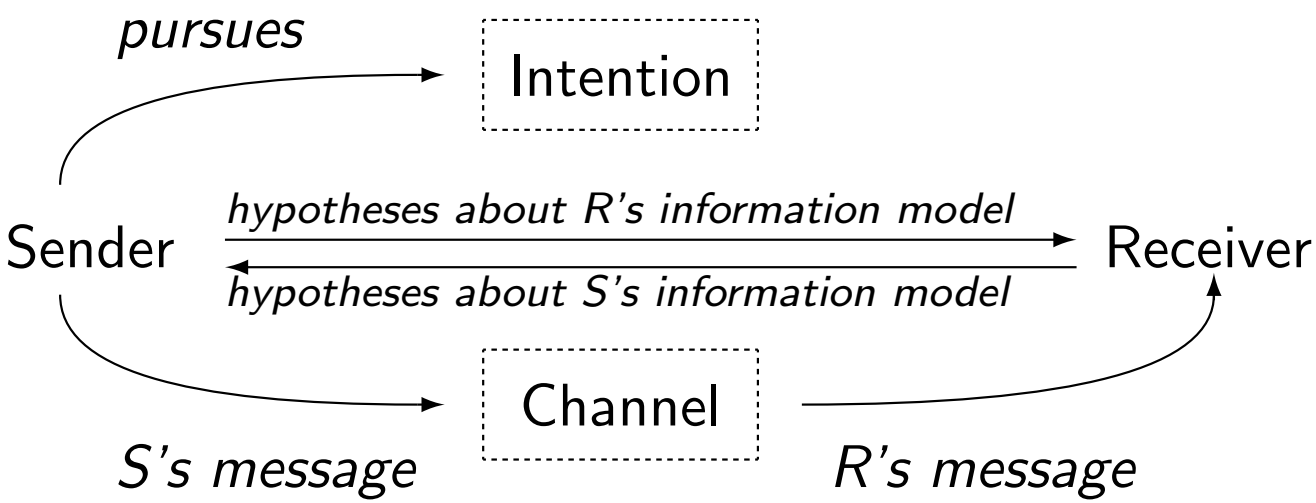
An interactional model of communication



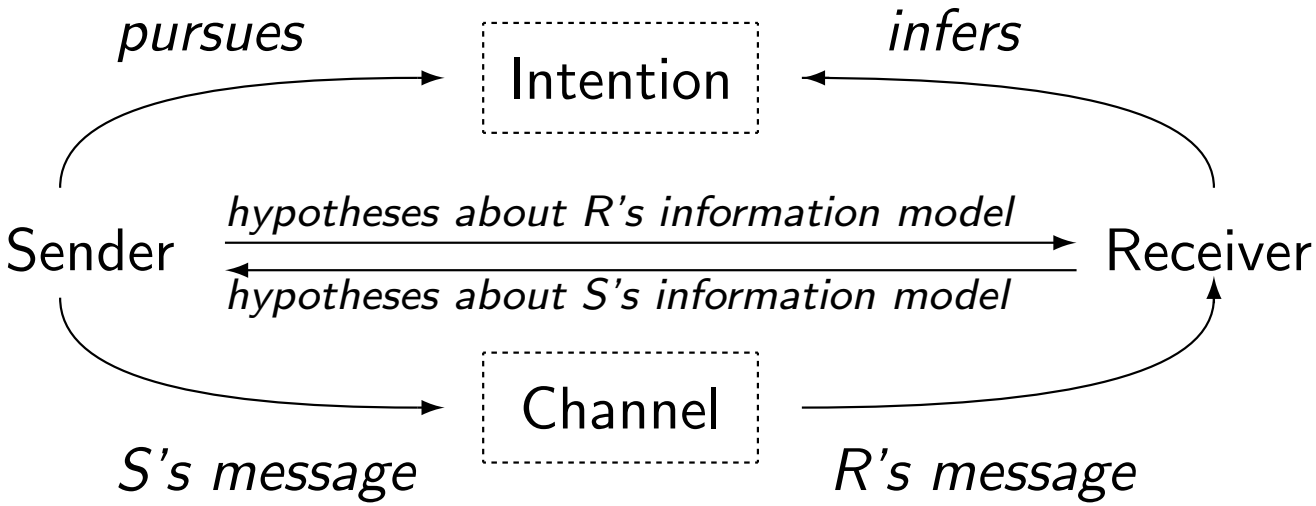
An interactional model of communication



An interactional model of communication



An interactional model of communication



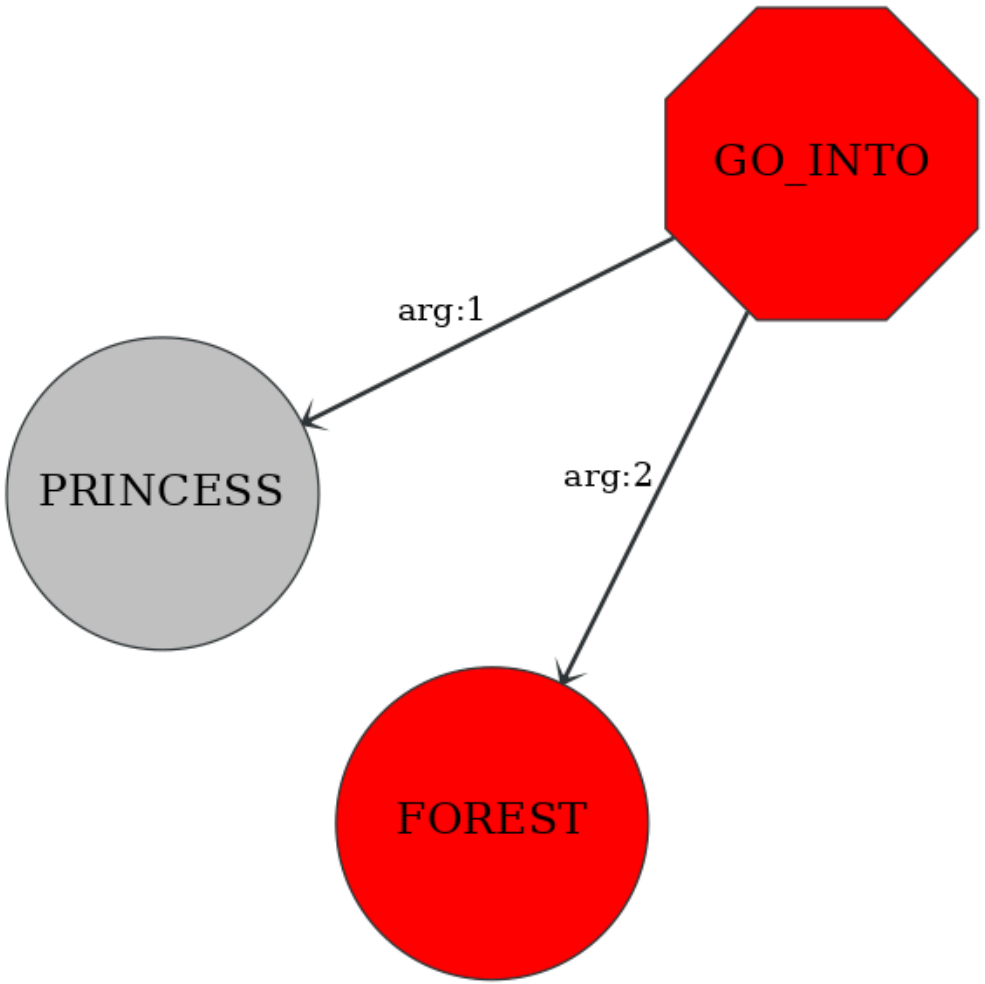
An example

Once upon a time there was a princess who went out into the forest and sat down at the edge of a cool well. She had a golden ball that was her favorite plaything. She threw it up high and caught it in the air . . .

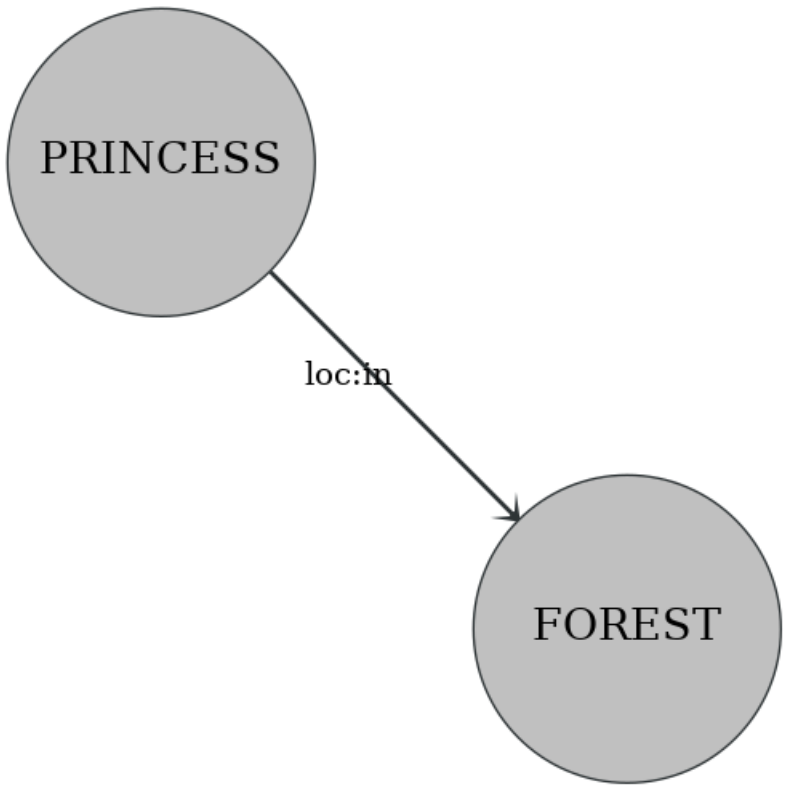
Once upon a time there was a princess . . . (input)

Once upon a time there was a princess . . . (proc.)

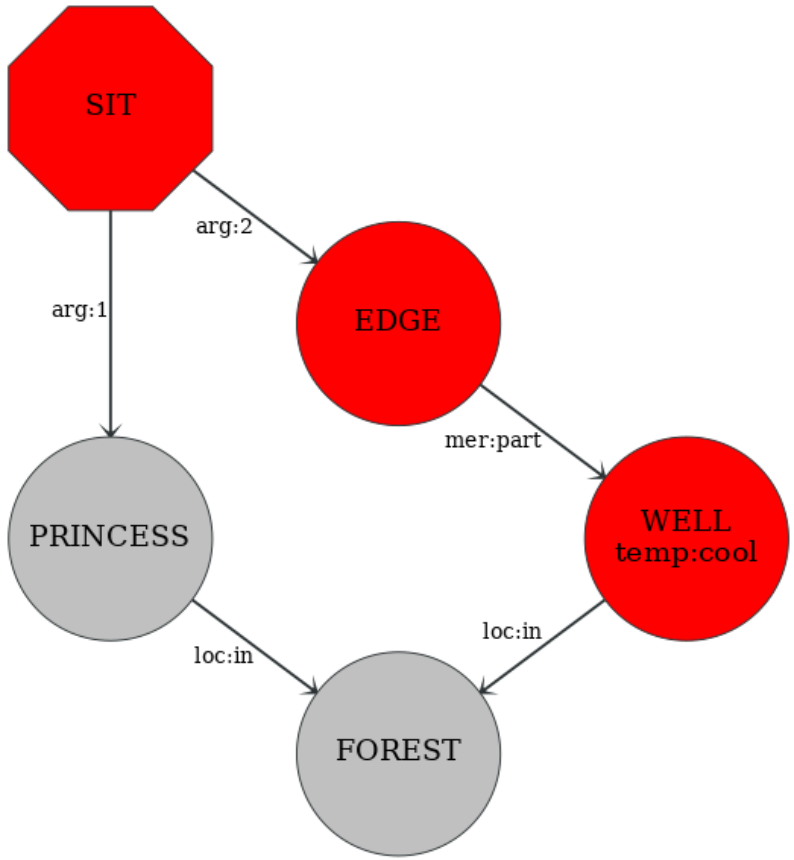
. . . who went out into the forest . . . (input)



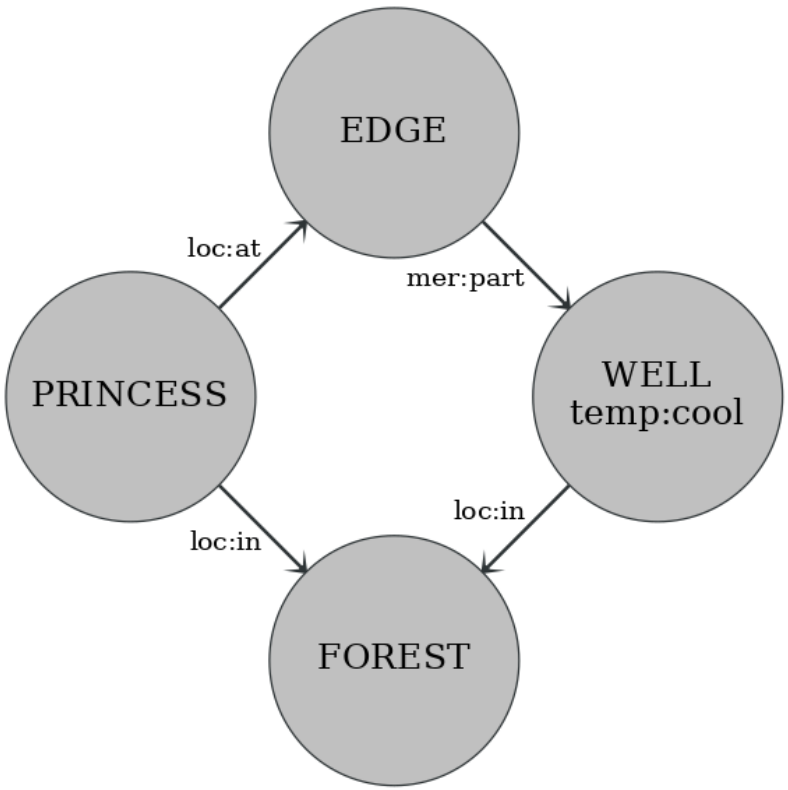
. . . who went out into the forest . . . (proc.)



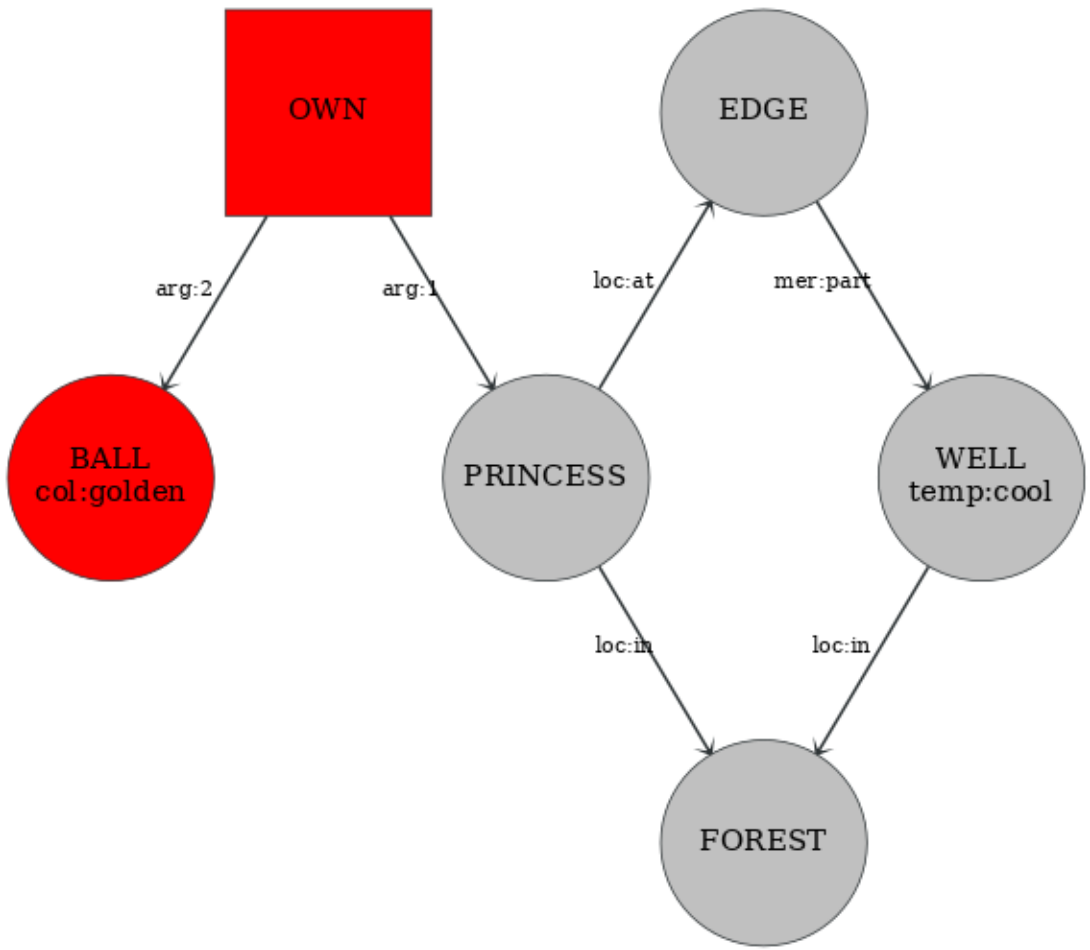
. . . and sat down at the edge of a cool well. (input)



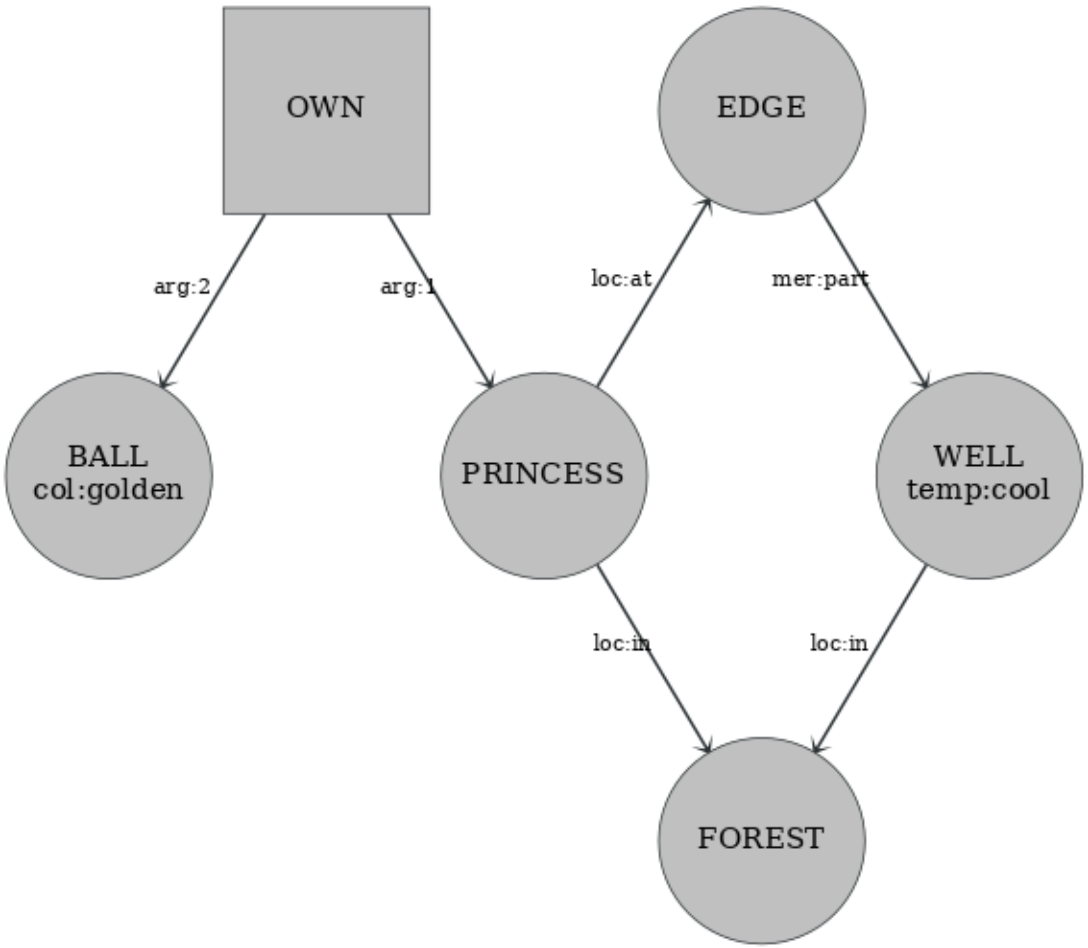
. . . and sat down at the edge of a cool well. (proc.)



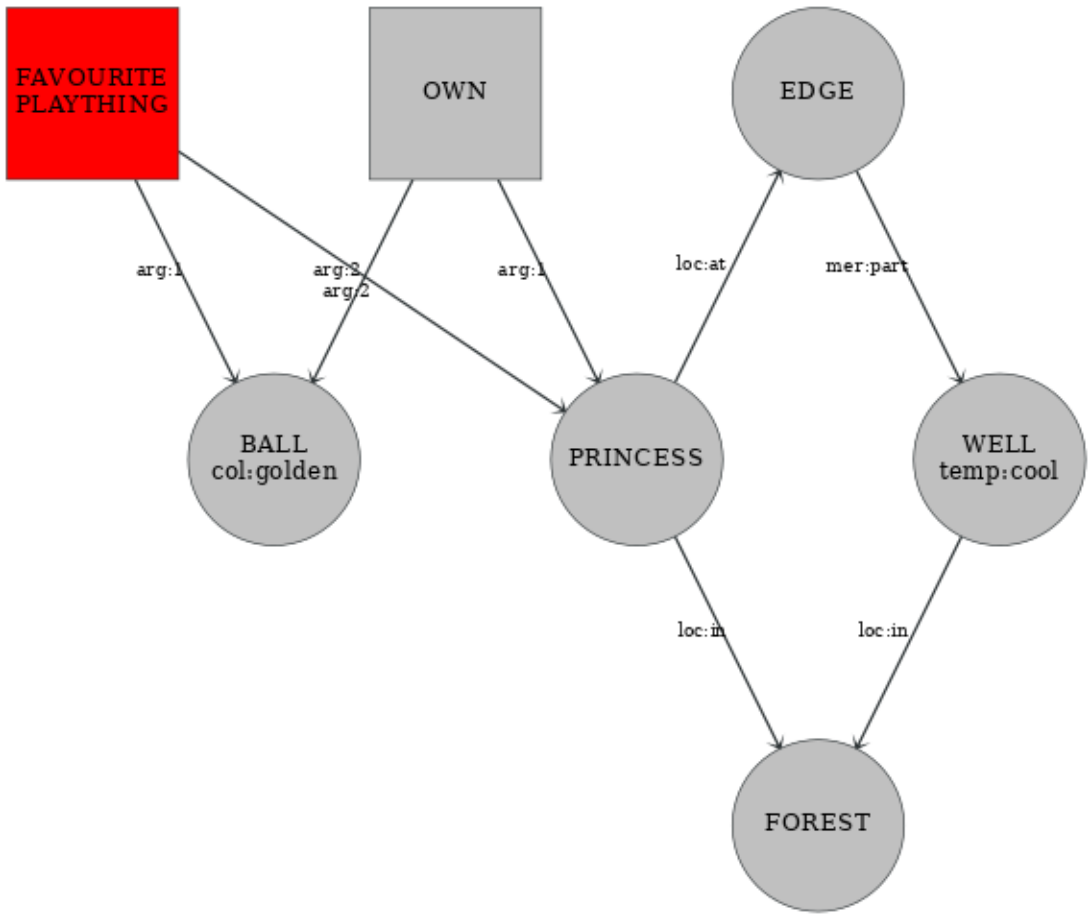
She had a golden ball . . . (input)



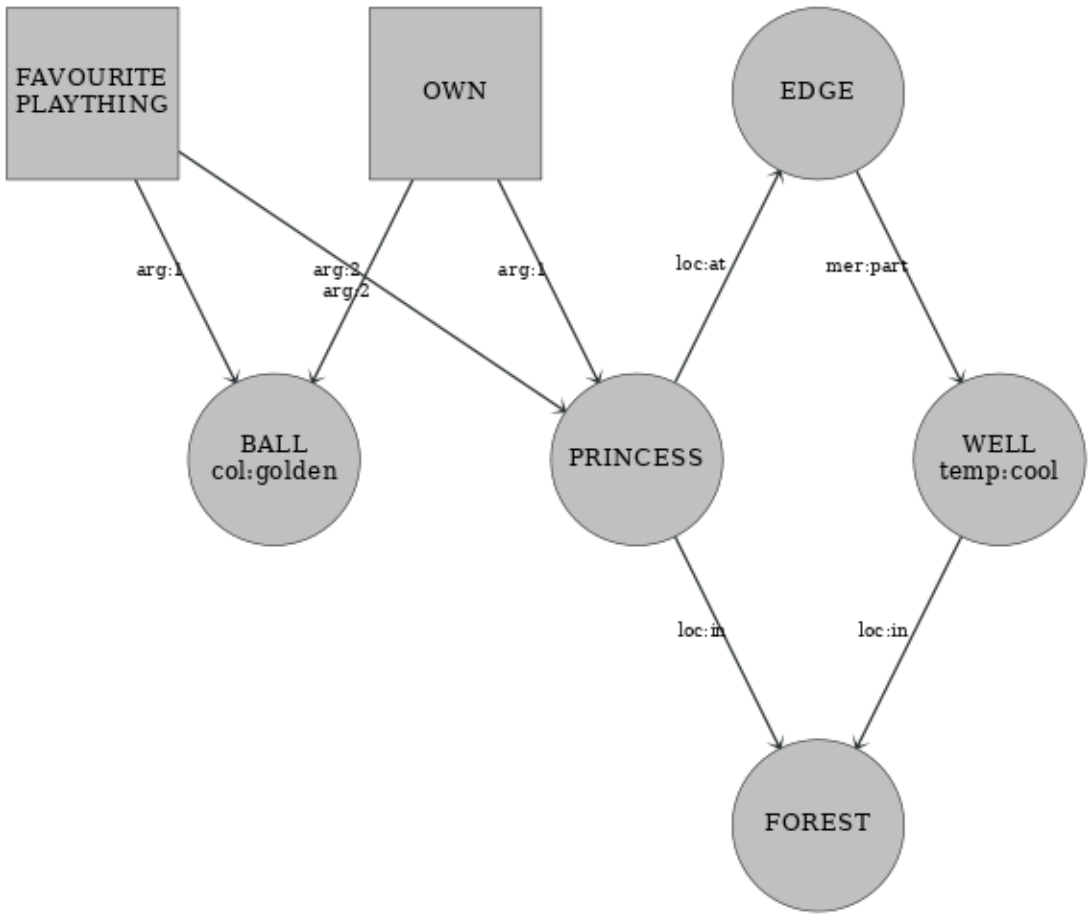
She had a golden ball . . . (proc.)



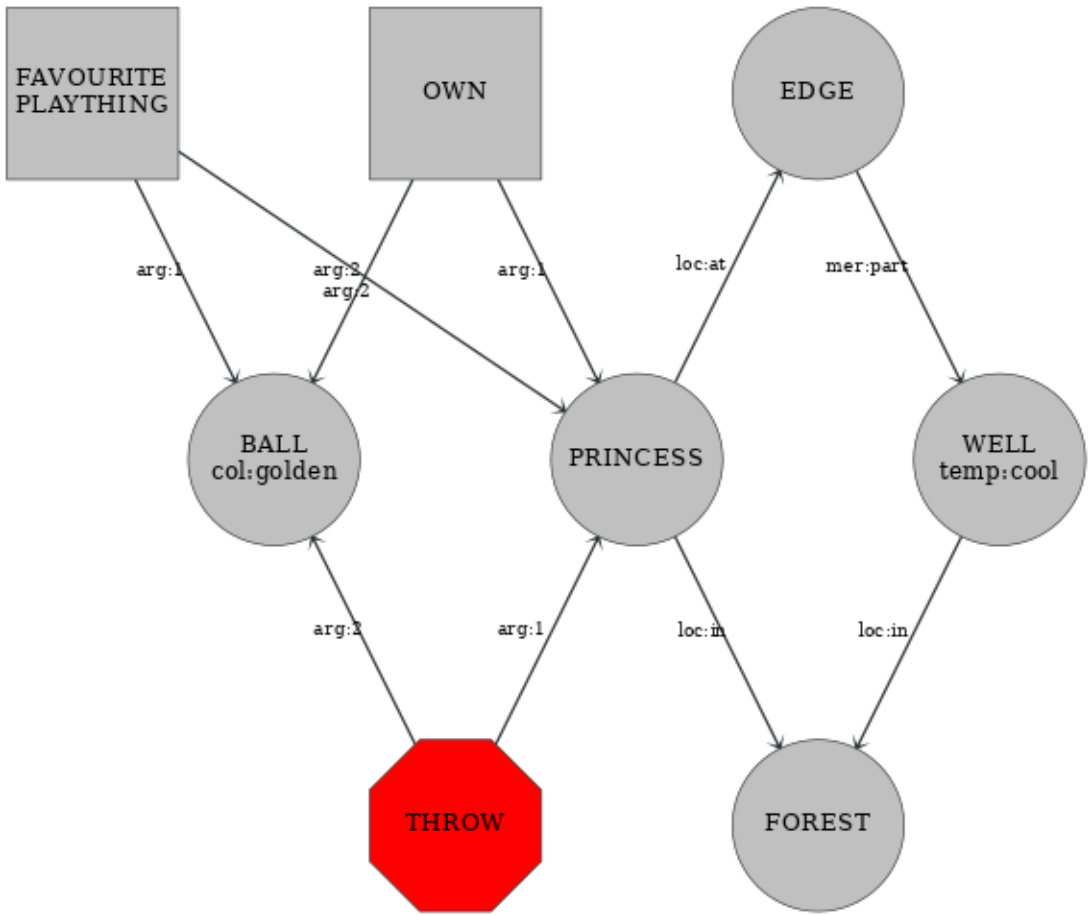
. . . that was her favourite plaything. (input)



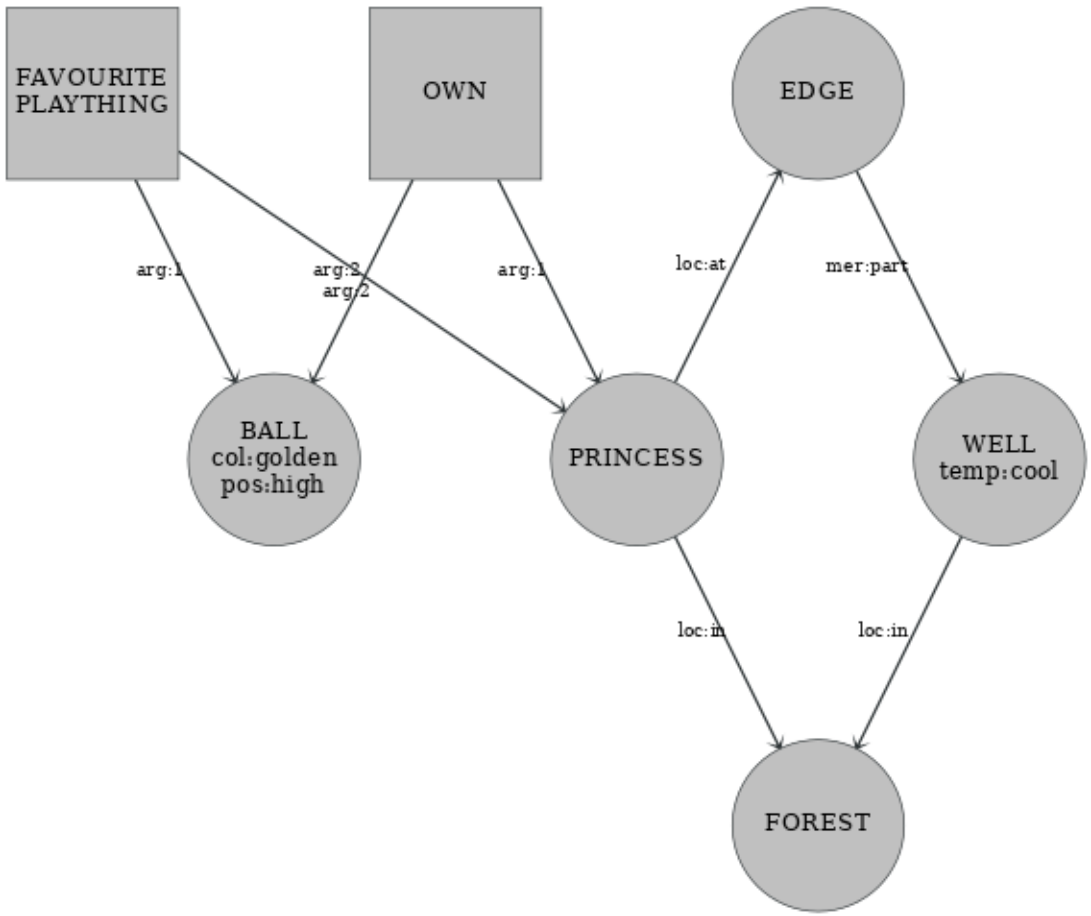
. . . that was her favourite plaything. (proc.)



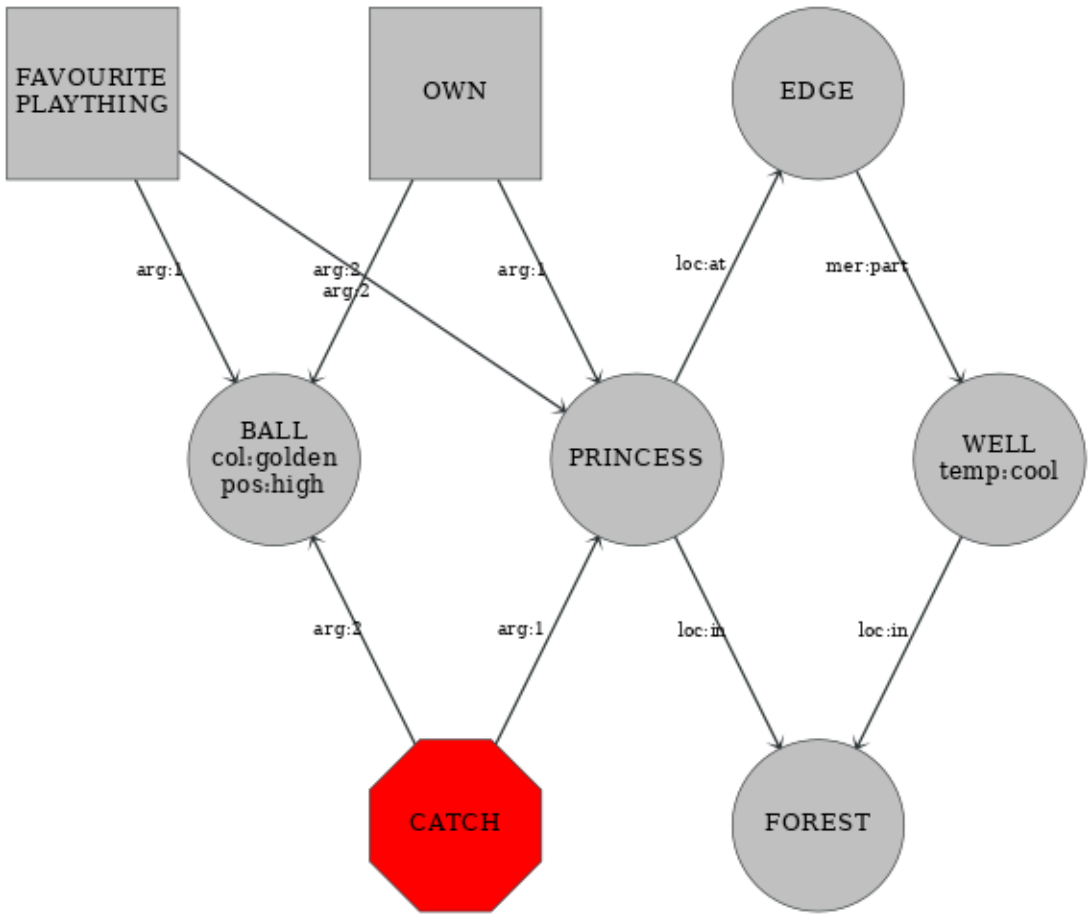
She threw it up high . . . (input)



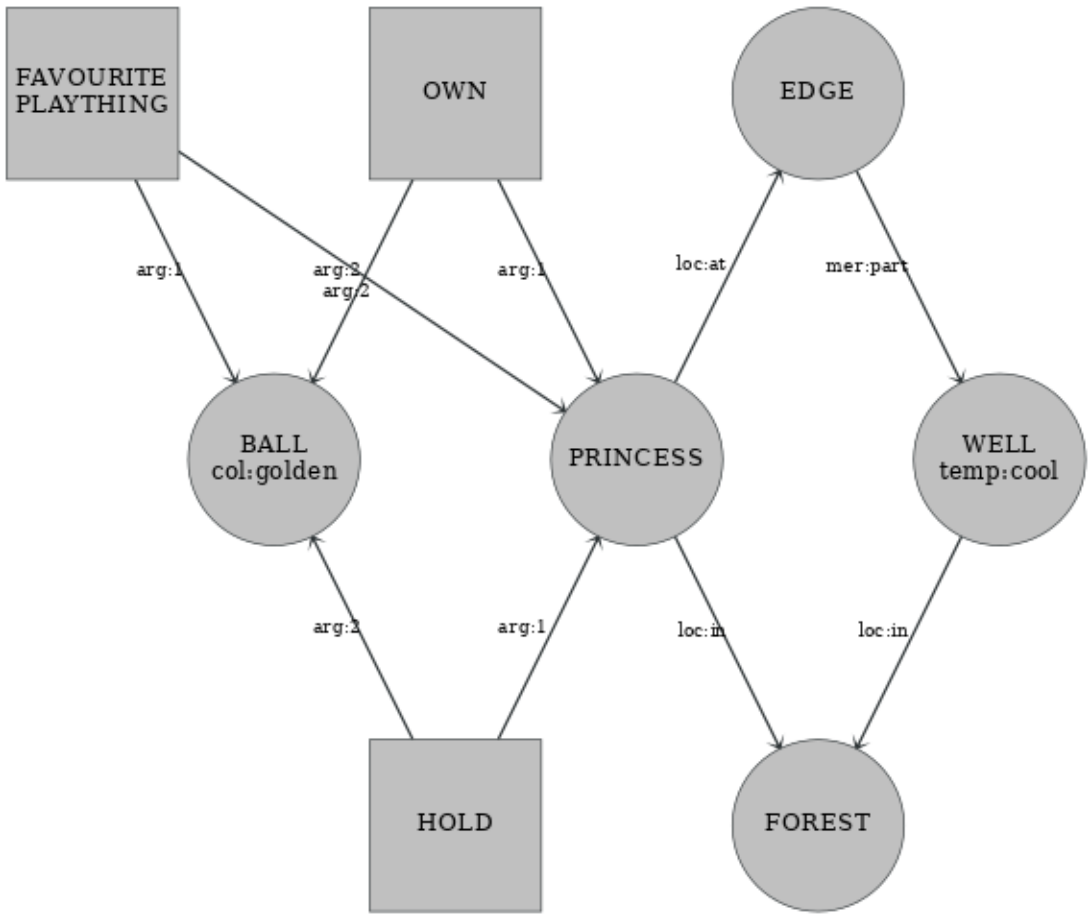
She threw it up high . . . (proc.)



. . . and caught it in the air . . . (input)



. . . and caught it in the air . . . (proc.)



The use of information models

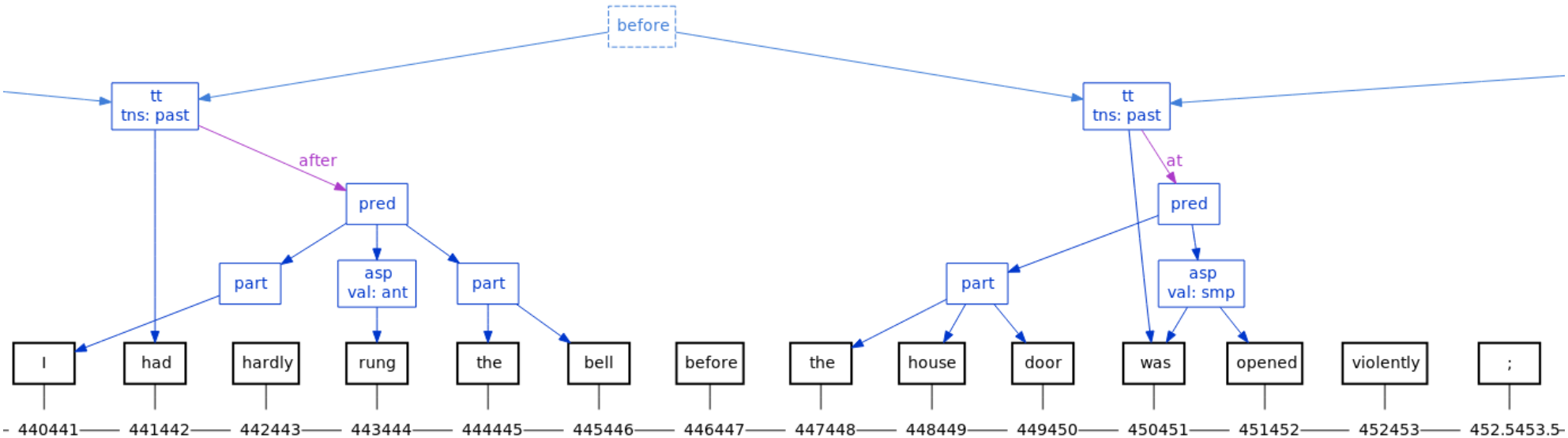
- Capture the dynamics of information transfer, i.e. communication, and its linguistic reflexes, e.g.
  - definiteness
  - conversational inferencing
- Ideally, information models can be extracted from texts through dependency parsing).

The temporal structure of an information model

- Information models have a temporal dimension; they consist of ‘slices’ representing one unit of narration (‘narrated time’).
- The order in which models are updated corresponds to the ‘narrating time’.

An annotation experiment (*The woman in white*)

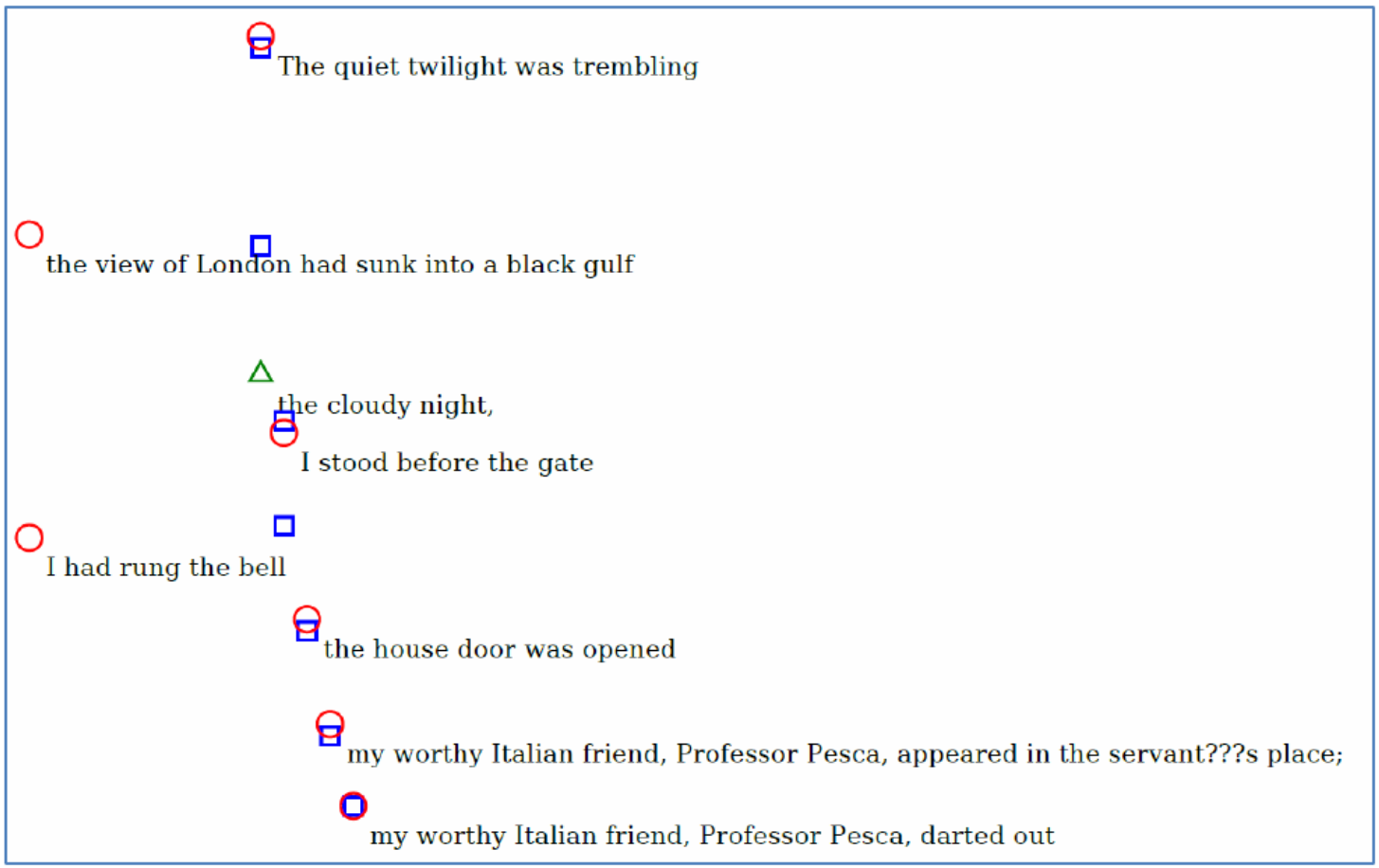
- Annotation scheme based on Klein (1994)[9,10]



[9] W. Klein (1994). *Time in Language*. London: Routledge.

[10] V. Gast et al. (2016). "Enriching TimeBank. Towards a more precise annotation of temporal relations in a text". In: *Proceedings of LREC, 2016*.

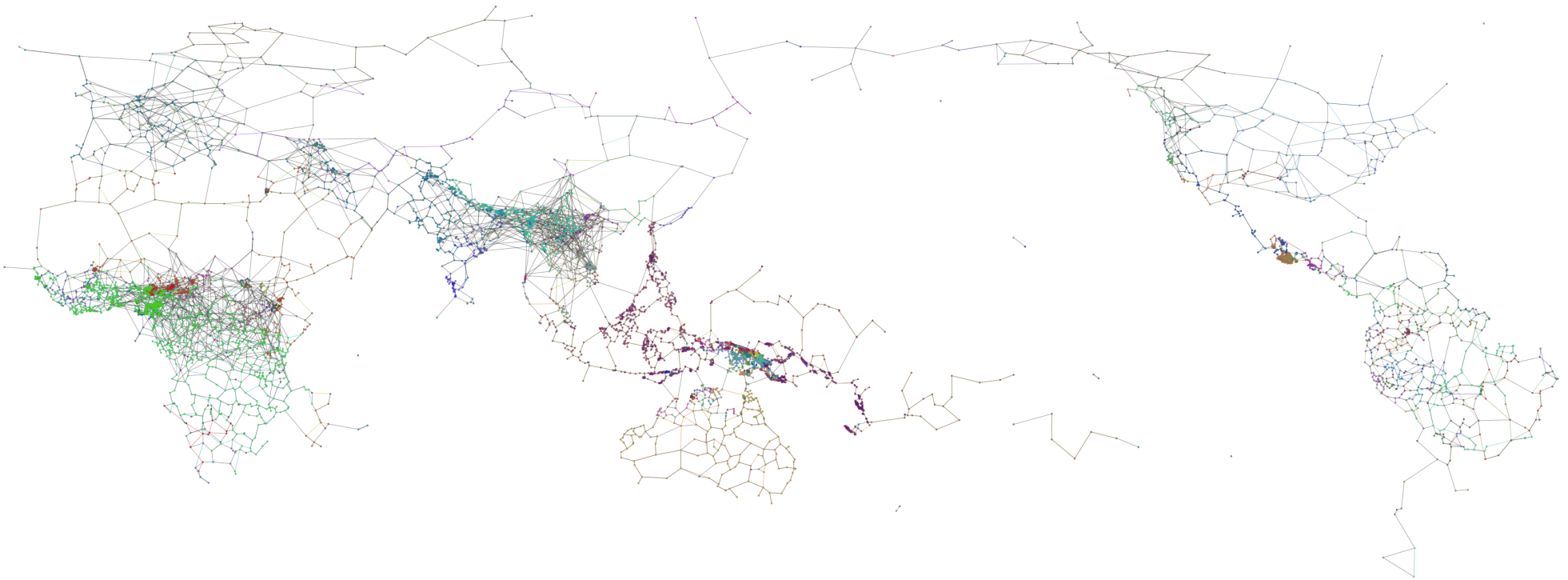
Temporal structure of (*The woman in white*)



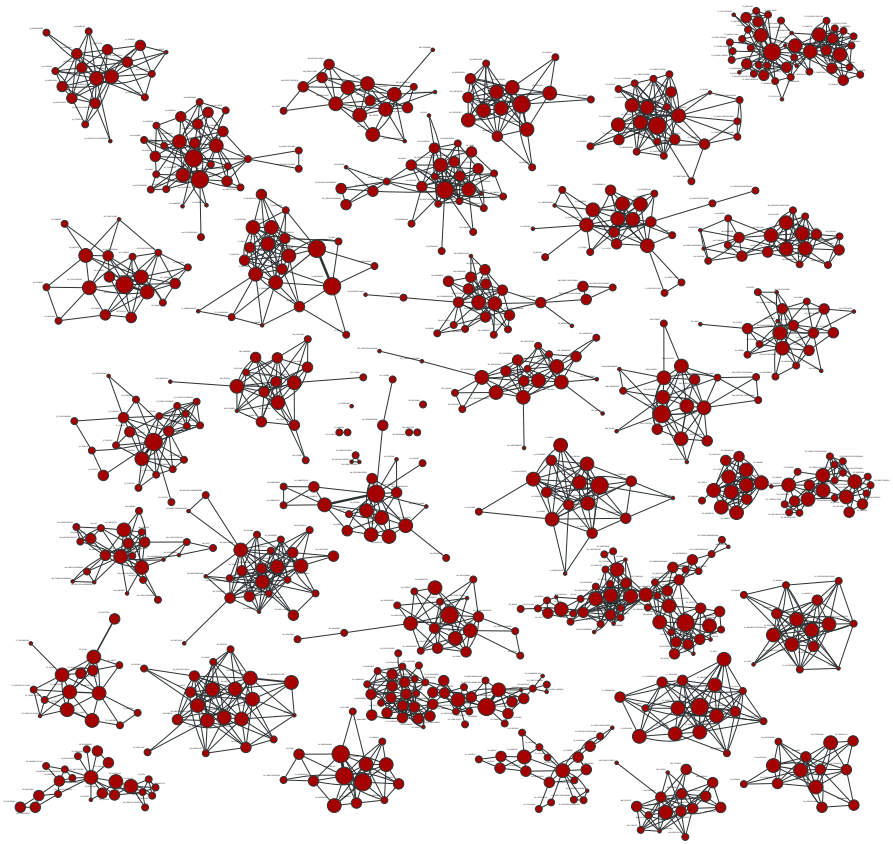
Summary

- Graphs are a tool of analysis with a lot of potential in many areas of linguistic analysis!
- Further uses:
  - Modelling the world's languages as a network ($\rightarrow$ understanding language contact)
  - Social networks in Shakespeare's plays
  - etc.

The world's languages



Social relations in Shakespeare's plays



Social relations in Shakespeare's plays

Software used

- Software used:
  - Python
    graph_tool: https://graph-tool.skewed.de/ (T. de Paula Peixoto)
  - R
    rgl: https://cran.r-project.org/web/packages/rgl (D. Adler, D. Murdoch)
    FactoMineR: http://factominer.free.fr/ (F. Husson, J. Josse, S. Lê)
  - Ruby
    GraphAnno: https://github.com/LBierkandt/graph-anno (L. Bierkandt)