The Tulip Project

a partial treatment of formal analogies

László Kálmán

Research Institute for Linguistics, HAS
Theoretical Linguistics Department, HAS/ELTE

kalman.laszlo@nytud.mta.hu

2019

plan

1. why “tulip”?
2. why “partial”?
3. why “treatment”?

history

- ancient use: ‘regularity’
- beginning of 19th century: exception to “sound laws”
- e.g., Hungarian *sirolm* $>$ *siralom* $\sim$ *siralm-* $\Rightarrow$ *sátor* $>$ *sátor* $\sim$ *sátr-*
- de Saussure: back to the ancient view, analogy is the force that makes linguistic systems coherent

analogy

example

$$A : B = C : ?$$

$$\begin{tikzcd} A \arrow[r] \arrow[d] & B \arrow[d] \\ C \arrow[r] & ? \end{tikzcd}$$

analogy

example

$$A : B = C : ?$$

$$\begin{tikzcd} \textit{son} \arrow[r] \arrow[d] & \textit{sons} \arrow[d] \\ \textit{sin} \arrow[r] & ? \end{tikzcd}$$

analogy

example

$$A : B = C : ?$$

$$\begin{tikzcd}
\textit{son} \arrow[r] \arrow[d] & \textit{sons} \arrow[d] \\
\textit{sin} \arrow[r] & \textit{sins}
\end{tikzcd}$$

Lepage (1996)

- transformation *son* $\rightsquigarrow$ *sons*
- transformation *son* $\rightsquigarrow$ *sin*
- composition of the two: *son* $\rightsquigarrow$ *sins*

motivation

Lepage was the first to offer a formal treatment of analogy, based on string operations (deletion and insertion, just like in the calculation of Levenshtein distances), and he has used it for finding regularities in large corpora.

However, this method is difficult to generalize to more sophisticated representations and operations; I felt that a more principled and abstract way of approaching analogy is called for.
In what follows, I will first show, starting from this simple string example, what analogy means at the most abstract level.

tulip

analogy

$$A : B = C : ?$$

$$\begin{tikzcd} A \arrow[r] \arrow[d] & B \arrow[d] \\ C \arrow[r] & ? \end{tikzcd}$$

tulip

analogy

$$A : B = C : ?$$

$$\begin{tikzcd} \textit{son} \arrow[r] \arrow[d] & \textit{son}\textcolor{red}{\textit{s}} \arrow[d] \\ \textit{sin} \arrow[r] & \textit{sin}\textcolor{red}{\textit{s}} \end{tikzcd}$$

tulip

analogy

$$A : B = C : ?$$

$$\begin{tikzcd}
\textit{son} \arrow[r] \arrow[d] & \textit{sons} \arrow[d] \\
\textit{sin} \arrow[r] & \textit{sins}
\end{tikzcd}$$

tulip

analogy

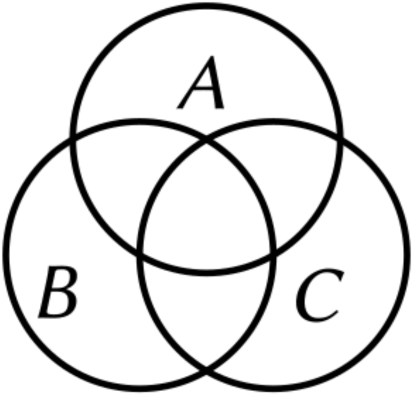
$$A : B = C : ?$$

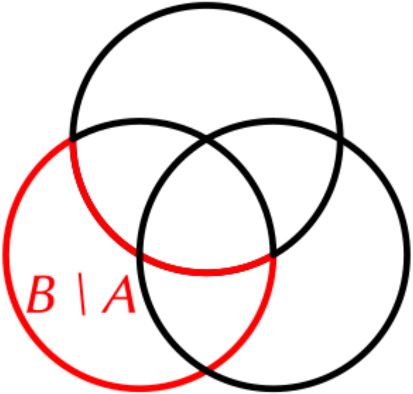
$$\begin{tikzcd}
\textit{son} \arrow[r] \arrow[d] & \textit{sons} \arrow[d] \\
\textit{sin} \arrow[r] & \textit{sins}
\end{tikzcd}$$

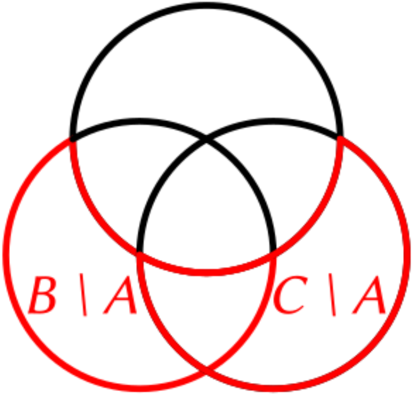
metaphorically

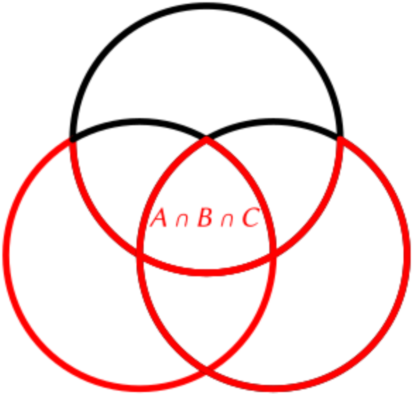
$$\textit{sons} \setminus \textit{son} = \textit{-s} \qquad \textit{sin} \setminus \textit{son} = \textit{-i-} \qquad \textit{son} \cap \textit{sons} \cap \textit{sin} = \textit{s- -n}$$

$$? = \textit{-s} \cup \textit{-i-} \cup \textit{s- -n} = \textit{sins}$$


A
B
C


$B \setminus A$


$B \setminus A$
$C \setminus A$


$A \cap B \cap C$

summary

It seems that the "tulip" is a suitable set theoretical metaphor, but how will we make it operational? What should be the members of our sets? They should be something like "ingredients" or "properties" of representations, but this is too general, and the calculi of such entities is often non-trivial.

why partial?

things i cannot do (and maybe do not want to)

- reduplication: *tami* : *tatami* = *paya* : ?
- metathesis: *top* : *pot* = *lead* : ?

things i cannot do (but would like to)

- sub- (auto-) segmental phenomena: *teeth* : *teethe* = *belief* : ?
- semantic phenomena: *sleep* : *slept* = *go* : ?
- true analogy is not based on individual examples, but legions of them, with varying frequencies

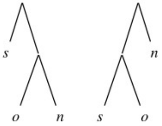
why treatment?

elements of a solution

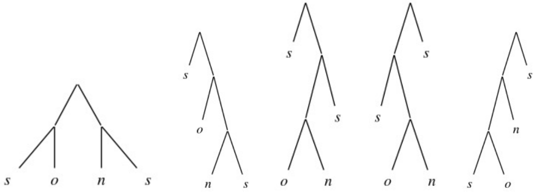
- approach 1: strings represented as sets of binary trees with ordered branches (“left” and “right”) with constraints as leaves
- approach 1: strings represented as partial orders over subsets of a set with an equivalence classification (corresponding to features)
- difference, intersection and “union” (combination) are defined accordingly

approach 1: trees

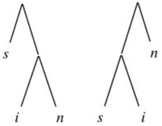
son



sons

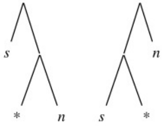


sin



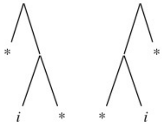
approach 1: trees

intersection

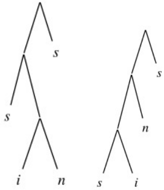


difference

difference



union



approach2: partial orders

brief description

- the universe $\mathcal{U} = \langle \mathcal{E}, \equiv \rangle$ (entities with an equivalence relation $\equiv \subseteq \mathcal{E}^2$, representing features)
- string constraints $\langle E, \leq \rangle$, with $E \subseteq \mathcal{E}$, a partial ordering $\leq$ over E
- for a simple string, $x_1^{(s)} \leq x_2^{(o)} \leq x_3^{(n)} \leq x_4^{(s)}$

the tricky part: operations

- they are defined over sets of constraints
- maps from "sons" to "son" and from "sin" to "son"
- difference: unmapped part; intersection: covered by both ranges
- linearization also yields sets of strings
- modelling sub- and autosegmental representations (partial order of timing points that start or end an autosegment)

Köszönöm a figyelmet!

references

Lepage, Yves and Shin-ichi, Ando, 1996. “Saussurian analogy: a theoretical account and its application”. *Proc. COLING ’96*, Vol. 2, pp. 717–722. ACL, Stroudsburg PA.