Gluing idioms back together:
A phraseo-combinatorial analysis

Manfred Sailer   Sascha Bargmann

July 28-29, 2021

Outline

Introduction

The **formal analysis of idioms**, like *kick the bucket* (≈ 'die') in (1)
or *pull strings* (≈ 'use connections') in (2),

  (1)   Our gold fish kicked the bucket last night.

  (2)   My boss pulled strings to get his current job.

has been oscillating between **two types of analysis:**

  A: **phrasal** analyses: focus on the unit-like character of idioms

  B: **combinatorial** analyses: focus on the autonomy of idiom parts

Goals of this talk:

- summarize the **main arguments for each type** of analysis
- propose an **analysis capturing the insights of both** types

Phrasal vs. combinatorial analyses

- Chomsky (1965):
  - **Idioms are** lexical units with internal structure, i.e. **phrase-like**.

- Wasow et al. (1983) and Nunberg et al. (1994):
  - **Idioms come in two versions:**
    - ⋆ Idiomatic **Phrase** (IP):
      semantically non-decomposable idiom analyzed as a fixed phrase
      coupled with its idiomatic meaning, e.g. *kick the bucket* (≈ 'die')
    - ⋆ Idiomatically **Combining** Expression (ICE):
      semantically decomposable idiom analyzed as consisting of two or more
      lexical entries that each contribute a part of the idiomatic meaning,
      e.g. *pull strings* (≈ 'use connections')

- Gazdar et al. (1985), Krenn & Erbach (1994), Sailer (2003), and
  Soehn (2009):
  - formalize the IP-vs.-ICE distinction in GPSG and HPSG.

Phrasal vs. combinatorial analyses

- Kay et al. (2015) and Bargmann & Sailer (2018):
  - extend the **combinatorial** analysis to non-decomposable idioms.

- Findlay (2019):
  - points out **two challenges for combinatorial approaches:**
    1. *collocational challenge***:** The idiomatic versions of the words need to be prevented from occurring independently of the idiom.
    2. *lexical explosion challenge***:** There is a new lexical entry for each word in each idiom.
  - suggests a **phrasal** analysis of idioms in a tree-grammar-based version of LFG.

Advantages of phrasal analyses

1. **Confinement of idiom parts** (no collocational challenge)**:**
   Only within an idiom's phrasal entry are the idiom's parts connected
   to an idiomatic meaning, so they cannot be used with this meaning
   outside of the entry.

2. **Lean lexicon** (no lexical explosion)**:**
   An idiom's parts are licensed directly through its phrasal entry, so
   there is no need for idiom-specific lexical entries for the idiom parts.

Disadvantages of phrasal analyses

1. It is **not clear how the degree of syntactic flexibility of an idiom is permitted** without recurring to diacritic marking of the possible syntactic configurations – as in Fraser (1970), Abeillé (1995).

2. Phrasal approaches do **not** seem to be the **appropriate** analytic tool **for some syntactic constellations**, see the next three slides.

Disadvantages of phrasal analyses

Idioms in relative clauses (RCs):

(3)   a. Parky pulled the strings$_k$ [$_{RC}$ that _____ $_k$ got me the job].
         (McCawley, 1981, 137)

      b. The strings$_k$ [$_{RC}$ that Pat pulled _____ $_k$] got Chris the job.
         (Nunberg et al., 1994, 510)

      c. John never pulled the strings$_k$ [$_{RC}$ that his mother told him
         should be pulled _____ $_k$]. (Henk v. Riemsdijk's example)

Licensing (Webelhuth et al., 2018):

- RC in (3a): no idiom parts → no licensing necessary

- RCs in (3b)–(3c): *pull* is licensed via the semantics of *strings$_k$*,
  which is present via the gap _____ $_k$

Disadvantages of phrasal analyses

Same idiom part occurring twice:

(4)   The beans have not been spilled yet, but will be spilled very soon.

Licensing:

but-clause in (4): *spill* is licensed via the semantics of *beans*,
which is the subject of both occurrences of *spill*.

Disadvantages of phrasal analyses

Pronominalization of idiom parts:

(5) a. Eventually she spilled all the $beans_k$. But it took her a few days to spill $them_k$ all. (Riehemann, 2001, 207)

  b. Kim's family pulled some $strings_k$ on her behalf, but $they_k$ weren't enough to get her the job. (Nunberg et al., 1994, 502)

Licensing (Webelhuth et al., 2018):

- *but*-clause in (5a): *spill* is licensed via the semantics of $beans_k$, which is present via the pronoun $them_k$

- *but*-clause in (5b): no idiom parts → no licensing necessary

Overview

factors ↓ // analyses →	phrasal	combin.
confinement of idiom parts	+	−
lean lexicon	+	−
idioms in relative clauses	−	+
same idiom part occurring twice	−	+
pronominalization of idiom parts	−	+

Intuition about idiomatic words

Riehemann (2001, 187):

> *The fact that idiomatic words cannot occur in their idiomatic meaning outside the idiom should be captured because these words do not have an existence independent of the idiomatic phrase they are a part of. By this I mean that there are no lexical entries for the individual idiomatic words, but only entries for the idiomatic phrases which contain them.*
>
> […]
>
> *Unfortunately, this intuition in its most natural form is not consistent with the current declarative logical frameworks, and probably could not be made consistent without additional machinery. Therefore I have worked out a formalization that works within the usual logical framework, although it does not capture the underlying intuition as straightforwardly.*

Basic ideas: WORDS set

WORDS-set on phrases: terminal nodes dominated by the phrase

(Riehemann, 2001, 184)

(193)

$$
\begin{bmatrix}
spill\_beans\_idiom\_phrase \ \& \ head\_subj\_phrase \\
\text{PHON } <diana\ spilled\ the\ beans> \\
\text{WORDS } \left\{ \boxed{1}, \boxed{2}, \boxed{3}, \boxed{4} \right\}
\end{bmatrix}
$$

$\boxed{1} \begin{bmatrix} diana \\ \text{PHON } <diana> \end{bmatrix}$

$$
\begin{bmatrix}
head\_comps\_phrase \\
\text{PHON } <spilled\ the\ beans> \\
\text{WORDS } \left\{ \boxed{2}, \boxed{3}, \boxed{4} \right\}
\end{bmatrix}
$$

$\boxed{2} \begin{bmatrix} i\_spill \\ \text{PHON } <spilled> \end{bmatrix}$

$$
\begin{bmatrix}
head\_spr\_phrase \\
\text{PHON } <the\ beans> \\
\text{WORDS } \left\{ \boxed{3}, \boxed{4} \right\}
\end{bmatrix}
$$

$\boxed{3} \begin{bmatrix} the \\ \text{PHON } <the> \end{bmatrix}$

$\boxed{4} \begin{bmatrix} i\_bean \\ \text{PHON } <beans> \end{bmatrix}$
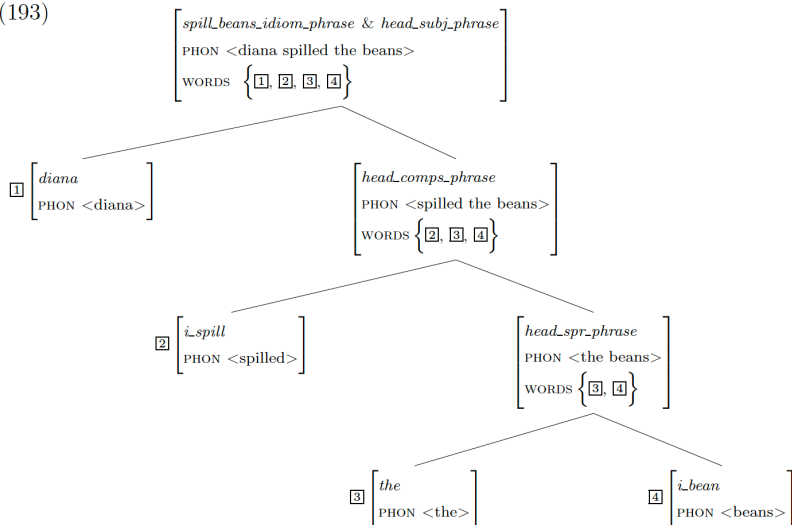
Riehemann's analysis of *spill beans*

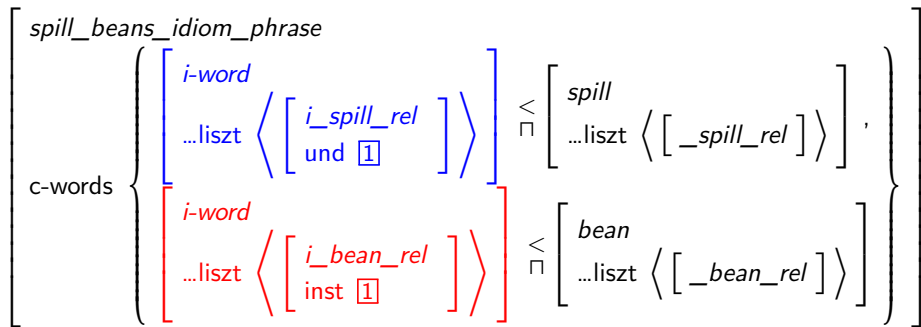- Phrasal type for idiomatic phrases: *spill_beans_idiom_phrase*.
- *spill_beans_idiom_phrase* has idiomatic words in its C-WORDS set.
- Idiomatic and literal words related by *skeptical default unification*
- No special lexical entry needed for idiomatic words!

$$
\begin{bmatrix}
\textit{spill\_beans\_idiom\_phrase} \\[2mm]
\text{c-words} \left\{
\begin{array}{l}
\begin{bmatrix} \textit{i-word} \\ \dots\text{liszt} \left\langle \begin{bmatrix} \textit{i\_spill\_rel} \\ \text{und } \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix}
\overset{<}{\sqcap}
\begin{bmatrix} \textit{spill} \\ \dots\text{liszt} \left\langle \begin{bmatrix} \textit{\_spill\_rel} \end{bmatrix} \right\rangle \end{bmatrix} , \\[6mm]
\begin{bmatrix} \textit{i-word} \\ \dots\text{liszt} \left\langle \begin{bmatrix} \textit{i\_bean\_rel} \\ \text{inst } \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix}
\overset{<}{\sqcap}
\begin{bmatrix} \textit{bean} \\ \dots\text{liszt} \left\langle \begin{bmatrix} \textit{\_bean\_rel} \end{bmatrix} \right\rangle \end{bmatrix}
\end{array}
\right\}
\end{bmatrix}
$$

Problems of Riehemann (2001)

- (Use of defaults)
- Full cross-classification of idiomatic phrases and clause types needed?
- WORDS-mechanism:
  - ▸ Motivation of the mechanism?
  - ▸ 1-to-1 relation between an idiomatic phrase and its idiomatic words.
- Problem with occurrence of idiom parts:

  (6) The beans were not spilled by Alex but will be spilled by Chris.

  (7) Eventually she spilled all the $beans_k$. But it took her a few days
       to spill $them_k$ all. (Riehemann, 2001, 207)

Basic ideas

Conserve the basic ideas of Riehemann (2001), but:

- Encode " $\underset{\sqcap}{<}$ " as a lexical rule.

- Replace the WORDS mechanism with a constraint on idiomatic phrases and a collocational restriction on idiomatic words.

- Completeness requirement of idioms is semantic (Webelhuth et al., 2018).

Idiomatic words as derived words

- Intuition: interpret Riehemann's "$\underset{\sqcap}{<}$" as lexical rule

- Object-level lexical rules à la Meurers (2001)

(8) a. $\begin{bmatrix} \textit{lexical-rule} \\ \text{in} \quad \textit{word} \\ \text{out} \quad \textit{word} \end{bmatrix}$

   b. $\textit{word} \longrightarrow$

   simple words: $\quad (L_1 \wedge \begin{bmatrix} \text{store} \ \langle\rangle \end{bmatrix}) \vee \ldots \vee (L_n \wedge \begin{bmatrix} \text{store} \ \langle\rangle \end{bmatrix}) \vee$

   derived words: $\quad \boxed{1} \begin{bmatrix} \text{store} \ \left\langle \begin{bmatrix} \textit{lex-rule} \\ \text{out} \ \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix}$

i-word-lexical-rule (version 1)

- Riehemann's "$\stackrel{<}{\sqcap}$" as (object-level) lexical rule *i-word-lexical-rule*
- Input: non-idiomatic version of a word
- Output: idiomatic version
- Input and output differ in their LID and CONT values.

$$
\begin{bmatrix}
\textit{i-word-lexical-rule} \\[1em]
\text{in} \quad \begin{bmatrix} \text{syns}|\text{loc} \begin{bmatrix} \text{cat} & \begin{bmatrix} \text{head}|\text{lid} \ \boxed{1} \end{bmatrix} \\ \text{cont} \ \boxed{2} \end{bmatrix} \end{bmatrix} \\[3em]
\text{out} \quad \begin{bmatrix} \text{syns}|\text{loc} \begin{bmatrix} \text{cat} & \begin{bmatrix} \text{head}|\text{lid} \ \boxed{3} \end{bmatrix} \\ \text{cont} \ \boxed{4} \end{bmatrix} \end{bmatrix}
\end{bmatrix}
\quad
\begin{array}{l}
\& \ \boxed{1} \neq \boxed{3} \\
\& \ \boxed{2} \neq \boxed{4}
\end{array}
$$
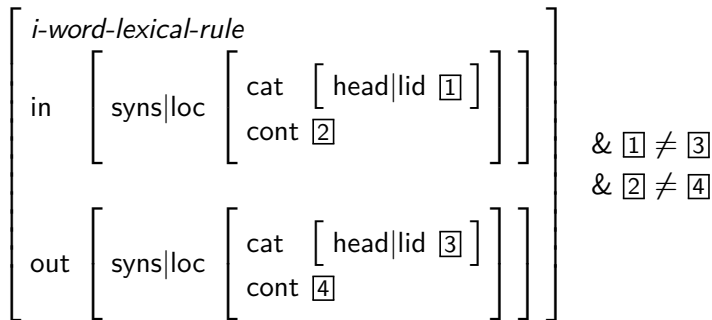
A phraseo-lexical entry of *spill beans*

- List-valued attribute C-WORDS on *phrase*
- Idioms are phrases with a non-empty C-WORDS list
- The (idiomatic) words constituting the idiom are on the C-WORDS list.

$$
\begin{bmatrix}
phrase \\
\text{c-words} \left\langle
\begin{bmatrix}
i\text{-}word\text{-}lexical\text{-}rule \\
\text{in} \ \begin{bmatrix} \text{lid} \ spill \end{bmatrix} \\
\text{out} \begin{bmatrix}
\text{lid} \quad spill\text{-}id \\
\text{cont} \begin{bmatrix} spill\text{-}id\text{-}rel \\ \text{und} \ \boxed{1} \end{bmatrix}
\end{bmatrix}
\end{bmatrix},
\begin{bmatrix}
i\text{-}word\text{-}lexical\text{-}rule \\
\text{in} \ \begin{bmatrix} \text{lid} \ bean \end{bmatrix} \\
\text{out} \begin{bmatrix}
\text{lid} \quad bean\text{-}id \\
\text{cont} \begin{bmatrix} bean\text{-}id\text{-}rel \\ \text{inst} \ \boxed{1} \end{bmatrix}
\end{bmatrix}
\end{bmatrix}
\right\rangle
\end{bmatrix}
$$
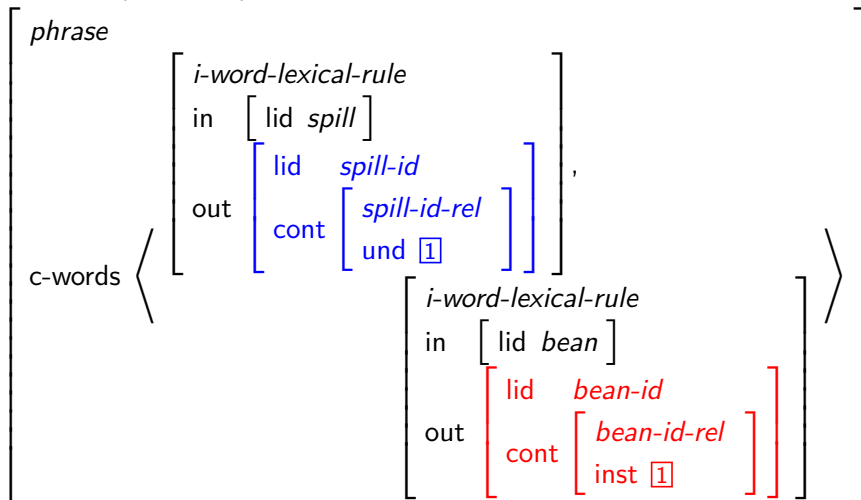
Constraint on idiomatic phrases

- List-valued attribute C-WORDS on *phrase*
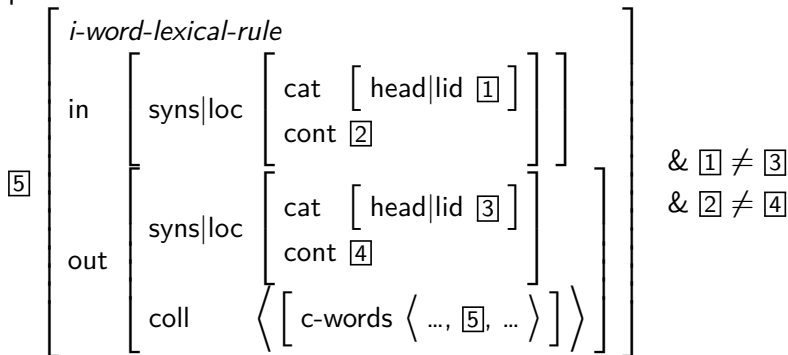- Idioms are phrases with a non-empty C-WORDS list
- Idiom-lexicon:

$$\begin{bmatrix} phrase \\ \text{c-words } nelist \end{bmatrix} \Rightarrow$$

$$\left(\begin{array}{l} \Big[\text{c-words } \Big\langle \big[\text{ in lid } spill \big], \big[\text{ in lid } bean \big] \Big\rangle \Big] \\ \vee \\ \Big[\text{c-words } \Big\langle \big[\text{ in lid } kick \big], \big[\text{ in lid } bucket \big] \Big\rangle \Big] \\ \vee \\ \dots \end{array} \right)$$

Collocations

- Distributional constraints on lexical items:
  - Bound words: Richter & Sailer (2003)
  - Negative polarity items: Richter & Soehn (2006)
  - External allomorphy (*a/an*): Soehn (2009)
- Prerequisite in (some) combinatorial approaches to idioms
- Simplest version:
  - List-valued feature COLL on lexical items
  - *sign* objects on the COLL list
  - A lexical item can only occur in a structure in which it is dominated by each of the elements on its COLL list.

i-word-lexical-rule (version 2)

The output of the *i-word-lexical-rule* is collocationally restricted to a phrase that has this instantiation of the lexical rule on its C-WORDS list.

$$
\boxed{5}
\begin{bmatrix}
\textit{i-word-lexical-rule} \\[4pt]
\text{in} \quad
\begin{bmatrix}
\text{syns|loc}
\begin{bmatrix}
\text{cat} \ \begin{bmatrix} \text{head|lid} \ \boxed{1} \end{bmatrix} \\
\text{cont} \ \boxed{2}
\end{bmatrix}
\end{bmatrix} \\[16pt]
\text{out}
\begin{bmatrix}
\text{syns|loc}
\begin{bmatrix}
\text{cat} \ \begin{bmatrix} \text{head|lid} \ \boxed{3} \end{bmatrix} \\
\text{cont} \ \boxed{4}
\end{bmatrix} \\[10pt]
\text{coll} \quad \left\langle \begin{bmatrix} \text{c-words} \ \left\langle \dots, \boxed{5}, \dots \right\rangle \end{bmatrix} \right\rangle
\end{bmatrix}
\end{bmatrix}
\quad
\begin{array}{l}
\& \ \boxed{1} \neq \boxed{3} \\
\& \ \boxed{2} \neq \boxed{4}
\end{array}
$$

Distributional constraint: simplified

- Idiomatic word triggers the presence of a licensing idiom (via COLL)
- Idiom must require the presence of the elements of its C-WORDS list:

  (9)   Constraint on idioms (first version):
        A phrase with a non-empty C-WORDS list must dominate the
        words identical with the OUT-values of the elements on its
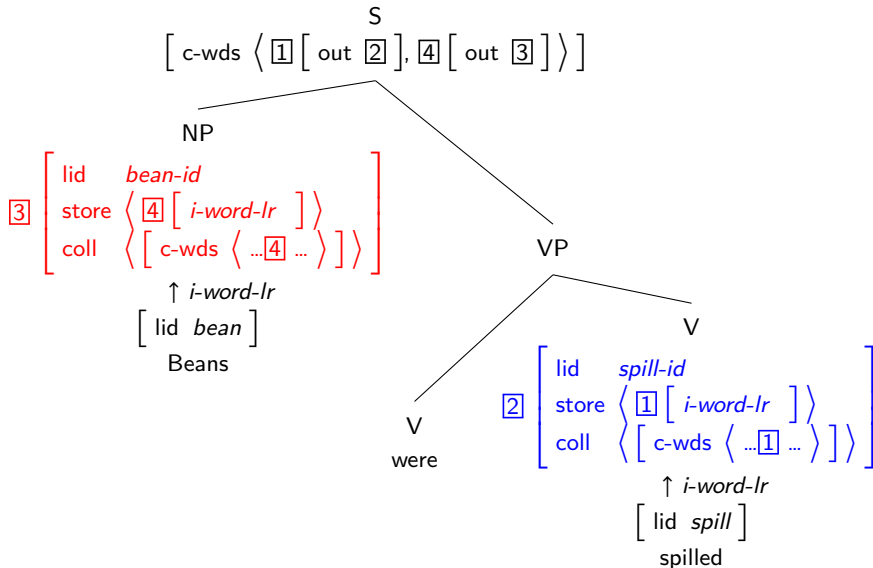   C-WORDS list.

(10) Syntactic flexibility:
      a. Alex spilled the beans.
      b. Beans were spilled.

(11) Co-occurrence requirement:
      $ Alex told me the beans.

Beans were spilled

Refining co-occurrence constraints 1

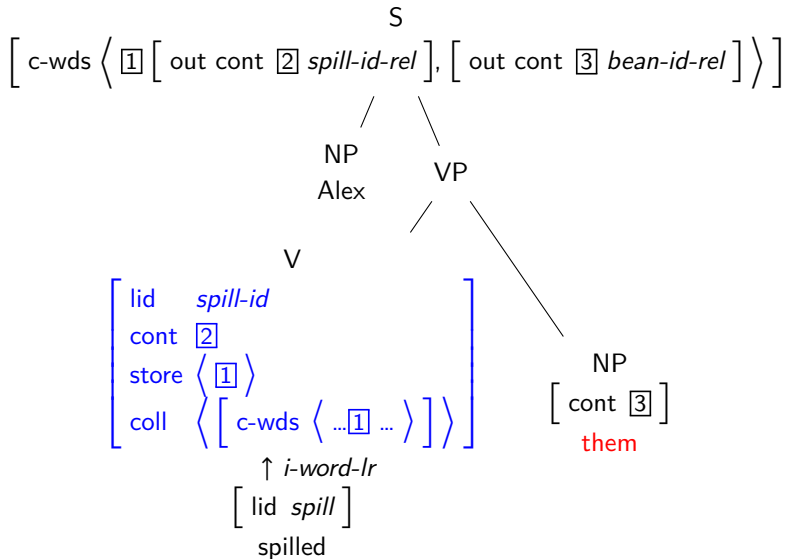- Pronominalized idiom part:

  (12)   Eventually she spilled all the beans$_k$. But it took her a few days
          to spill them$_k$ all. (Riehemann, 2001, 207)

- Problem: *spill* requires the occurrence of the idiomatic phrase, but
  there is no idiomatic *beans*

- Solution (Webelhuth et al., 2018): The idiomatic phrase only requires
  the occurrence of the content!

(13)   Constraint on idioms (final version):

        For each phrase *p* and for each object *o* on *p*'s C-WORDS list,
        p dominates a sign whose CONT value is identical with *o*'s
        OUT|…|CONT value.

Alex spilled them.

$$\left[\, \text{c-wds} \left\langle\, \boxed{1} \left[\, \text{out cont} \boxed{2} \textit{ spill-id-rel} \,\right], \left[\, \text{out cont} \boxed{3} \textit{ bean-id-rel} \,\right] \right\rangle \right]$$

S

NP
Alex

VP

V

$$\begin{bmatrix} \text{lid} & \textit{spill-id} \\ \text{cont} & \boxed{2} \\ \text{store} & \left\langle\, \boxed{1} \,\right\rangle \\ \text{coll} & \left\langle\, \left[\, \text{c-wds} \left\langle\, ...\boxed{1}... \,\right\rangle \,\right] \,\right\rangle \end{bmatrix}$$

↑ *i-word-lr*

$$\left[\, \text{lid } \textit{spill} \,\right]$$

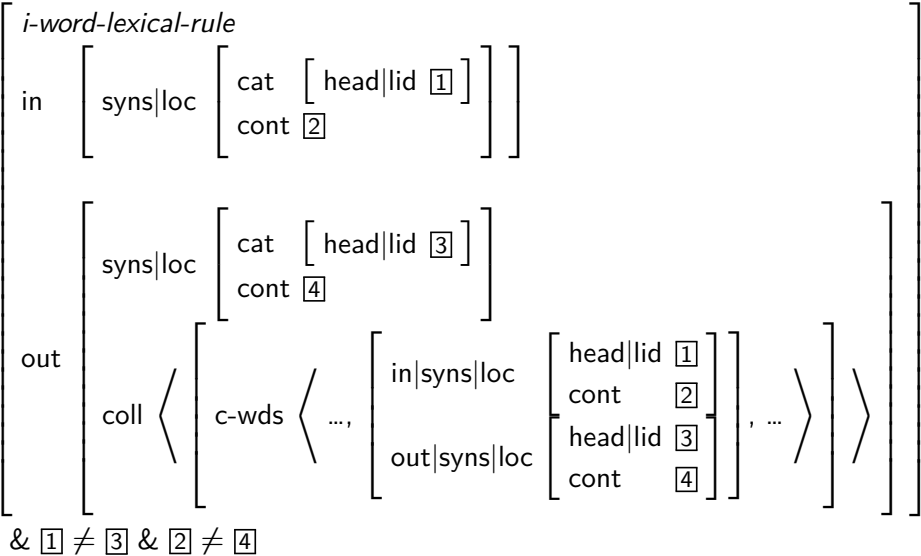spilled

NP

$$\left[\, \text{cont} \boxed{3} \,\right]$$

them

# Refining co-occurrence constraints 2

- One of the idiomatic words occurs twice.

  (14) The beans weren't spilled by Alex but will be spilled by Chris.

- But: There is only one corresponding *i-word-lex-rule* object on the phrase's C-WORDS list.
- Solution: Refined COLL requirement of idiomatic words:
  - occurrence of an *i-word-lexical-rule* object
  - with the same CONTENT/LID value on the output

i-word-lexical-rule (final version)

$$
\begin{bmatrix}
\textit{i-word-lexical-rule} \\[4pt]
\text{in} \begin{bmatrix} \text{syns|loc} \begin{bmatrix} \text{cat} \begin{bmatrix} \text{head|lid} \;\boxed{1} \end{bmatrix} \\ \text{cont} \;\boxed{2} \end{bmatrix} \end{bmatrix} \\[20pt]
\text{out} \begin{bmatrix}
\text{syns|loc} \begin{bmatrix} \text{cat} \begin{bmatrix} \text{head|lid} \;\boxed{3} \end{bmatrix} \\ \text{cont} \;\boxed{4} \end{bmatrix} \\[10pt]
\text{coll} \left\langle \begin{bmatrix} \text{c-wds} \left\langle \ldots, \begin{bmatrix} \text{in|syns|loc} \begin{bmatrix} \text{head|lid} \;\boxed{1} \\ \text{cont} \quad\;\; \boxed{2} \end{bmatrix} \\ \text{out|syns|loc} \begin{bmatrix} \text{head|lid} \;\boxed{3} \\ \text{cont} \quad\;\; \boxed{4} \end{bmatrix} \end{bmatrix}, \ldots \right\rangle \end{bmatrix} \right\rangle
\end{bmatrix}
\end{bmatrix}
$$

& $\boxed{1} \neq \boxed{3}$ & $\boxed{2} \neq \boxed{4}$

The beans were not spilled by A. but will be spilled by C.
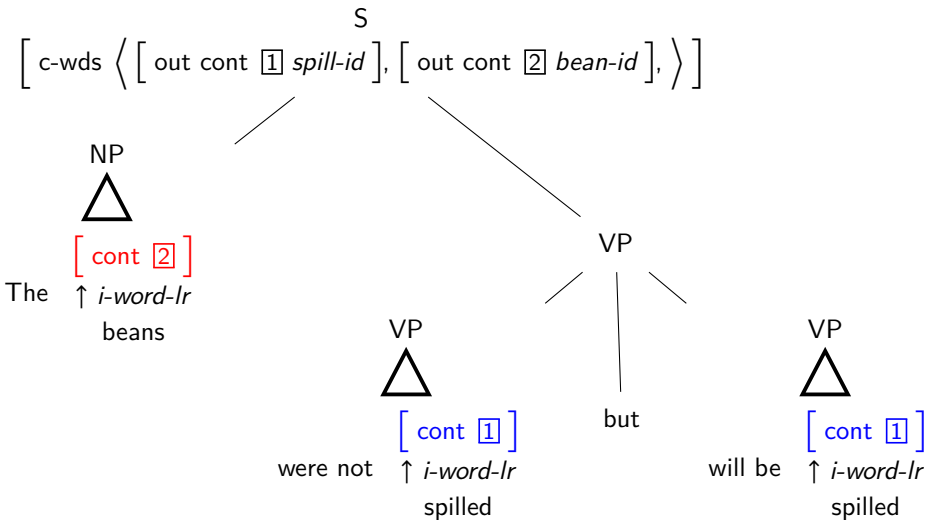
Summary of the analysis

- Shared with phrasal analyses:
  - ▸ Confinement of idiom component: idiomatic words do not exist outside the idiom
  - ▸ Lean lexicon: one general *i-word-lexical-rule*; a single specification for each idiom in the constraint on phrases with $\left[\text{ c-words } \textit{nelist} \right]$.
- Shared with combinatorial analyses:
  - ▸ Idioms in relative clauses: Constraint on idiom only requires the presence of idiomatic content, not of idiomatic words.
  - ▸ Same idiom part occurring twice: No 1-to-1 requirement on C-WORDS elements and dominated idiomatic words.
  - ▸ Pronominalization of idiom parts: Semantic requirement

Summary of the analysis

factors ↓ // analyses →	phrasal	combin.	**phraseo-combin.**
confinement of idiom parts	+	−	+
lean lexicon	+	−	+
idioms in relative clauses	−	+	+
same idiom part occurring twice	−	+	+
pronominalization of idiom parts	−	+	+

Summary

- Phrasal analysis with all the flexibility of a combinatorial analysis
- Combinatorial analysis without lexical explosion
- Idioms are represented as single unit
- Integrates Webelhuth et al.'s (2018) collocational approach based on semantic representations.
- Captures Riehemann's (2001) intuition, using independently motivated tools:
  - object-level lexical rules: standard in HPSG
  - collocation module: assumed in combinatorial approaches
- However: Not transferable to Kay et al. (2015): address the collocational challenge by extended selection rather than collocations
- Gain in coverage beyond what combinatorial approaches can do already?

Thank you!

References I

Abeillé, Anne. 1995. The flexibility of French idioms: A representation with Lexical MTree Adjoining Grammar. In Martin Everaert, Eric-Jan van der Linden, André Schenk & Ron Schreuder (eds.), *Idioms. structural and psychological perspectives*, 15–42. Hillsdale: Lawrence Erlbaum Associates.

Fraser, Bruce. 1970. Idioms within a transformational grammar. *Foundations of Language* 6(1). 22–42.

Kay, Paul, Ivan A. Sag & Dan Flickinger. 2015. A lexical theory of phrasal idioms. Manuscript.
www1.icsi.berkeley.edu/~kay/idiom-pdflatex.11-13-15.pdf.

McCawley, James D. 1981. The syntax and semantics of English relative clauses. *Lingua* 53. 99–149.

References II

Meurers, Walt Detmar. 2001. On expressing lexical generalizations in HPSG. *Nordic Journal of Linguistics* 24(2). 161–217. doi:10.1080/033258601753358605. http://purl.org/dm/papers/lexical-generalizations.html. Special issue on 'The Lexicon in Linguistic Theory'.

Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70. 491–538.

Richter, Frank & Manfred Sailer. 2003. Cranberry words in formal grammar. In Claire Beyssade, Olivier Bonami, Patricia Cabredo Hofherr & Francis Corblin (eds.), *Empirical issues in formal syntax and semantics*, vol. 4, 155–171. Paris: Presses Universitaires de Paris-Sorbonne.

References III

Richter, Frank & Jan-Philipp Soehn. 2006. *Braucht niemanden zu scheren*: A survey of npi licensing in german. In Stefan Müller (ed.), *Proceedings of the 13th international conference on head-driven phrase structure grammar, varna 2006*, 421–440. `http://cslipublications.stanford.edu/HPSG/2006/richter-soehn.pdf`.

Riehemann, Susanne Z. 2001. *A constructional approach to idioms and word formation*: Stanford University dissertation.

Soehn, Jan-Philipp. 2009. Lexical licensing in formal grammar. `http://nbn-resolving.de/urn:nbn:de:bsz:21-opus-42035`.

Webelhuth, Gert, Sascha Bargmann & Christopher Götze. 2018. Idioms as evidence for the proper analysis of relative clauses. In Manfred Krifka, Rainer Ludwig & Mathias Schenner (eds.), *Reconstruction effects in relative clauses*, 225–262. Berlin and Boston: de Gruyter.