Focus, Givenness and Information Status
Annotating Corpora with Information Structure
ESSLLI 2014

Kordula De Kuthy and Arndt Riester

August 19, 2014
Questions to be addressed

- What is focus?
- What constitutes a focus theory?
- How can we describe the influence of focus on the meaning of sentences, and on their appropriateness in discourses?
- How can we detect focus in corpus data?
What is focus?

Answers provided in the literature:

- Focus is the **answer to a question** (to an explicit but also to an implicit one).
- Focus is the **informative** part of an utterance.
- Focus is the part of an utterance that signals **alternatives**.
- Focus indicates **new, or important / contrastive** information.
- Focus is **asserted / at issue**.
- Focus is often signalled by **prosodic or syntactic prominence**. *(language-dependent)*
Two important theoretical contributions in the past

(1) a. Who is laughing?
   b. $\text{JOHN} \text{focus} [\text{is laughing}]_{\text{given/background}}$.

- Of the two most influential focus frameworks in the past 30 years, one concentrates on the focus part, the other on the given part.
- Mats Rooth’s Alternative Semantics (Rooth 1985, 1992, 1996, 2010) is based on the idea that focus triggers (contrastive) alternatives.
- Roger Schwarzschild (Schwarzschild 1999) develops a technical givenness notion.
- Contemporary theories of information structure, such as Büring (2008); Beaver & Clark (2008); Wagner (2012) and others, mainly build on, and combine, ideas from Rooth and Schwarzschild.
Schwarzschild (1999): GIVENness, AvoidF and other constraints on the placement of accent
Prerequisite: focus projection (discussed yesterday)

- *Focus* is not equivalent with the word that carries the *pitch accent*.
- In West-Germanic languages (English, Dutch, German), focus originates in an accented word, and then *projects* onto larger phrases (Gussenhoven 1983, 1992, 1999; Rochemont 1986; Selkirk 1984, 1995; Winkler 1997).

(2) \[ \text{[MARY} f [bought} f [a} f [\text{book} f [\text{about} f \text{BATS} f]} f ] f ] f ] F \]
Limits of focus projection – I

- Focus projection is not mandatory.
- Normally, focus projects in order to achieve discourse congruence.
- But focus projection cannot solve all problems.
- Focus does not project from the head of a phrase to an argument.

(3) \[ \text{Mary} \_f [\text{bought}_f [\text{a}_f [\text{BOOK}_f [\text{about bats}]]_f]_f]_f]_F. \]

- In sentence (3) *about bats* cannot get an F-mark.
- This can become problematic.
Givenness principle (Schwarzschild 1999)

- Constituents which are not F-marked, must be given.
- Reversal: constituents which are not given must be F-marked.
- They must either be accented or “borrow” an F-mark by means of focus projection.

(14) \[ \text{[M Ary}_f \text{[bought}_f \text{[a}_f \text{[BOOK}_f \text{[about bats]}_f}_f}_f]_F. \]

- I.e. (3) can only be used in a context which already talks about bats. Otherwise it will be infelicitous.
- Caution: the givenness principle does not imply that given constituents must not be F-marked.

(15) a. Do you prefer vanilla or walnut?
    b. I prefer WALnut\(_F\). (given but F-marked)
Limits of focus projection – II

Modifiers are typically adjuncts (optional information), not arguments. F-marks on modifiers do not project to the head noun.

(16) {What did John drive?}

a. #He drove a BLUE\_F convertible.
   \textit{Givenness principle violated, because the F cannot project onto “convertible”, which is new.}

b. ✓ He drove [the\_f [convertible\_f [of\_f [his\_f MUM\_f]\_f]\_f]\_f]\_F.  
   (“convertible” receives f via horizontal projection)
Schwarzschild’s goal

Provide a unified theory that accounts for the accent patterns in the following cases:

(17) a. Why don’t you have some French toast?  
b. I’ve forgotten how to MAKE French toast.  \(\text{(newness)}\)

(18) a. John’s mother voted for Bill.  
b. No, she voted for JOHN.  \(\text{(contrast / correction)}\)

(19) a. Who did John’s mother vote for?  
b. She voted for JOHN.  \(\text{(question-answer)}\)

- Halliday (1967) redefines \textit{newness} to capture all these cases. This is not intuitive.
- Schwarzschild: a unified theory should not make reference to new information at all.
- What we need to do is redefine \textit{GIVENness}.
Phonological observations on English

- Prominence indicates novelty.  
  Wrong!

- Lack of prominence indicates givenness.  
  Correct!
AVOIDF

- GIVENNESS: Constituents which are not F-marked must be given.
- In order to avoid a violation of GIVENNESS, we might simply F-mark everything, e.g. place a pitch accent everywhere.
- But this is not what is happening.
- There must be an additional constraint which tells us to use accents sparingly: AVOIDF

AVOIDF:

"F-mark as little as possible, without violating GIVENNESS."
Question-answer congruence (Halliday 1967)

- An appropriate answer to a wh-question must have F-marking on the constituent corresponding to the wh-phrase.

(20)  a. What did Mary do?
     b. She [praised$_f$ [her$_F$ BROTHER$_f$]$_f$]$_F$.

(21)  a. What did John’s mother do?
     b. She [[PRAISED$_f$ him]$_F$.]

(22)  a. Who did John’s mother praise?
     b. She praised HIM$_F$. 
Consequences of Schwarzschild’s approach

Recall:

(23) What did John drive?
    #He drove a BLUE\textsubscript{F} convertible.
    
    ▶ GIVENNESS violated.

(24) John drove Mary’s red convertible. What did he drive before that?
    ✓ He drove a BLUE\textsubscript{F} convertible.
    
    ▶ No GIVENNESS violation
    ▶ Question-answer congruence is lost on Schwarzschild’s account.
What does *given* mean after all?

(25) John drove Mary’s red convertible. What did he drive before that?

a. He drove [a BLUE\(F\) convertible].

- The phrase [a BLUE\(F\) convertible] is not F-marked itself. Hence it must be *given*. But is it?
- The indefinite phrase introduces a new referent into the discourse (Heim 1982; Kamp & Reyle 1993).
- Also intuitively, since it contains new material, the phrase is not entirely given.
- The same goes for the entire sentence.
GIVEN (Schwarzschild 1999)

An utterance $U$ counts as GIVEN iff it has a salient antecedent $A$ and

a. if $U$ is type $e$, then $A$ and $U$ co-refer;
b. otherwise: modulo $\exists$-type shifting, $A$ entails the $\exists$-F-closure of $U$. 
What does that mean?

► If an expression is of type $e$ (e.g. *he*), it must be co-referential, e.g. with the earlier mentioned *Paul*.

► If an expression is of type $\langle \alpha, \beta \rangle$ (e.g. the verb *moves*), it must be entailed by some other expression in the discourse context (e.g. *walks*).

► How can we say that an arbitrary expression entails another one?
Existential Typeshift

\[
[\text{walks}] = \lambda x. \text{walk}(x)_{<e,t>} \quad [\text{moves}] = \lambda x. \text{move}(x)_{<e,t>}
\]

- Type shift to proposition level: replace lambdas by existential quantifiers.

\[
\exists x. \text{walk}(x)_t \quad \exists x. \text{move}(x)_t
\]

- Check whether the typeshifted antecedent entails the typeshifted “anaphor”.
- If such an entailment relation can be established then \textit{moves} is GIVEN.
Existential Typeshift

\[ \text{[walks]} = \lambda x. \text{walk}(x)_{(e,t)} \quad \text{[moves]} = \lambda x. \text{move}(x)_{(e,t)} \]

- Type shift to proposition level: replace lambdas by existential quantifiers.
  \[ \exists x. \text{walk}(x)_t \models \exists x. \text{move}(x)_t \checkmark \]

- Check whether the typeshifted antecedent entails the typeshifted “anaphor”.
- If such an entailment relation can be established then \text{moves} is GIVEN.
Existential F-Closure

- F-marks function as “wildcards” in the entailment process.
- The phrase *a BLUE$_F$ convertible* should count as GIVEN if it occurred after *a red convertible*.
- **∃-F-closure**: replace F-marked part of a constituent by an existentially bound variable (*an X-colored convertible*).
- Perform existential typeshift (here: from quantifier to proposition).

\[
\begin{align*}
\text{a red convertible} & \models \text{an X-colored convertible} \\
\exists P \exists x [\text{conv}(x) \land \text{red}(x) \land P(x)] & \models \exists P \exists X \exists x [\text{conv}(x) \land X(x) \land P(x)]
\end{align*}
\]
Predicting focus and accent in an OT-model

- Schwarzschild’s model only allows for indirect predictions of focus and accent placement.
- A set of candidates with different F-distributions is generated.
- Each candidate is tested for its compliance with GIVENNESS.
- AVOIDF: If several candidates pass GIVENNESS, the one with the least F-marks is chosen.
A very simple example

(26) Whom did John₁’s mother₂ praise?
    She₂ praised him₁.

Candidates:

i. She praised HIM₉.
ii. She \([\text{praised}_f \text{HIM}_f]_F\).
iii. She PRAISED₉ him.
iv. She \([\text{PRAISED}_f \text{him}]_F\).
v. SHE₉ praised him.
vi. SHE₉ praised HIM₉.
vii. . . .
Checking for GIVENness

(26)  Whom did John₁’s mother₂ praise?

  i.  She₂ praised [HIM₁]ₚ.

\[
\begin{align*}
\left[\text{John’s mother}_2\right] & \leftrightarrow \left[\text{she}_2\right] \checkmark \\
\left[\text{did } \ldots \text{ praise}\right] & \models \left[\text{praised}\right] \checkmark \\
\left[\text{John}_1\right] & \leftrightarrow \left[\text{him}_1\right] \checkmark
\end{align*}
\]

- \left[\text{praised } \text{him}_1\right] is not GIVEN but F-closure saves the day.

\[
\left[\text{did } \ldots \text{ praise whom}\right] \models \left[\text{praised HIM}_F\right] \checkmark
\]

- Hence, candidate (i) is good (and minimally F-marked).
Other candidates?

- iii. She $\text{PRAISED}_F$ him.
- (iii) is also minimally F-marked, but...
- $[\text{PRAISED}_F \text{him}_1]$ is not GIVEN.
- Even after applying F-closure (replacing the verb by “did something”) the context does not entail that “somebody did something to John”.
- Hence, (iii) is not a good candidate.
Convertible example

(27)  a. John drove Mary’s red convertible. What did he drive before that?
     b. He drove her BLUEF convertible.

By the same reasoning, we can show that all constituents of (27b) are GIVEN.
Upshot

- On Schwarzschild’s account, question-answer congruence is lost.
- Computations for natural data can become extremely complex.
- Summary of the account: If F-marks are distributed in an appropriate manner, then every constituent is – technically – GIVEN.
- Very unintuitive givenness notion
- The account may get almost all examples right, but it is not really useful for annotation purposes.
- In the following, we will develop another approach to givenness that follows a different tradition but integrates some of Schwarzschild’s ideas.
- First, back to the basics of givenness...
Information status
Information status of referring expressions

- Information status – originally – describes the **degree of givenness / salience / cognitive activation / accessibility** of referring expressions.
- Goal: distinguish classes of expressions in text or spoken discourse in a way that is as **fine-grained** as possible and still **reproducible** by non-experts with **high reliability**
- Prince (1981) distinguishes **textually / situationally evoked, inferrable and new** expressions.
- Prince (1992) introduces two dimensions (discourse status vs. hearer status):

<table>
<thead>
<tr>
<th>hearer-old</th>
<th>hearer-new</th>
</tr>
</thead>
<tbody>
<tr>
<td>discourse-old</td>
<td>given, old, active</td>
</tr>
<tr>
<td>discourse-new</td>
<td>unused, familiar, known</td>
</tr>
</tbody>
</table>
Information status

- Cognitive activation: Chafe (1994) distinguishes highly salient (active / given / consciously available), less salient (semi-active / accessible / unconscious), and non-salient (inactive / new)

- Other notable classifications (each one with their own use of terminology) are Gundel et al. (1993) *(givenness hierarchy)*, Lambrecht (1994); Poesio & Vieira (1998); Eckert & Strube (2000); Ariel (2001) *(accessibility theory)*, Nissim et al. (2004); Götze et al. (2007); Riester et al. (2010)

- An overview and partial comparison provided in Baumann & Riester (2012)
Our system

- *RefLex* scheme for the annotation of information structure (Baumann & Riester 2012; Riester & Baumann ms.)
- Goal: combine the approach by Schwarzschild (1999) with earlier accounts of information status
- Address a number of problems of earlier accounts
- Enable annotations on natural data that are both reliable and fine-grained
- Two levels:
  1. Referential Information Status (*referring expressions*)
  2. Lexical Information Status (*non-referring expressions*)
The analysis of referring expressions has a long history in linguistics and philosophy.

Two early claims by Frege (1891, 1892) on definite descriptions and proper nouns / names:

1. The felicitous use of a definite presupposes the existence of an entity to which the definite can refer.
2. This presupposed entity is unique, i.e. there is exactly one entity that satisfies the description.

This is indeed the case for certain definites: the sun, the present Pope, Gottlob Frege, the President of the United States, the positive square root of 4, . . .
Analysis of referring expressions (cont.)

- Russell (1905): Quantificational analysis of definite descriptions
- Strawson (1950): Criticism of Russell, restoring a variant of Frege’s referential account
- There is something wrong about the uniqueness assumption in definites like *the table, the cup, the man, the ant, the molecule, . . .*
- Uniqueness must be relativized to different *types of contexts.*
### Types of context

Definite descriptions can be unique with respect to different context types:

<table>
<thead>
<tr>
<th>Context</th>
<th>Label (RefLex)</th>
<th>Phenomenon</th>
</tr>
</thead>
<tbody>
<tr>
<td>previous discourse context</td>
<td>\textit{r-given}</td>
<td>coreference</td>
</tr>
<tr>
<td></td>
<td></td>
<td>anaphora</td>
</tr>
<tr>
<td>communicative situation</td>
<td>\textit{r-given-sit}</td>
<td>symbolic deixis</td>
</tr>
<tr>
<td></td>
<td>\textit{r-environment}</td>
<td>gestural deixis</td>
</tr>
<tr>
<td>frame / scenario</td>
<td>\textit{r-bridging}</td>
<td>bridging /</td>
</tr>
<tr>
<td></td>
<td></td>
<td>associative anaphora</td>
</tr>
<tr>
<td>following discourse context</td>
<td>\textit{r-cataphor}</td>
<td>cataphora</td>
</tr>
<tr>
<td>global context</td>
<td>\textit{r-unused}</td>
<td>global uniqueness</td>
</tr>
</tbody>
</table>

An (indefinite) expression, which refers non-uniquely, receives the label \textit{r-new}. 

Annotation conventions for referential expressions

- Annotate all referring expressions (i.e. what is called DP in generative linguistics, or NP in e.g. computational linguistics):

  (28) a cat, she, his, the table, this ugly lamp, John, Chancellor Merkel, Eddie’s, Tübingen, someone, freedom, squirrels, the guy who is sleeping etc.

- Quantified DPs:

  (29) every participant, many suitcases, a lot of work, few factories etc.

- In case a preposition is present, it is included in the markable:

  (30) (asked) for the bill, (went) to Tunisia, because of the new law, with several friends etc.
Annotation conventions (cont.)

- Appositions are included:

  (31) John Smith, the ambassador; a drink, which turned out to be mango lassi; Harry, who hasn’t been seen for two weeks

- Focus-sensitive particles are *not* included in the markable:

  (32) *(only) a snack; (even) Helga; the assignment*, too
r-given (old, active, textually evoked)

Coreference, uniqueness in previous discourse

(33) I met a friend yesterday.
   a. [He] told me a story. (pronoun)
   b. [The friend] came from Hamburg. (same noun)
   c. [The old chap] was very tired. (different expression)
   d. I hadn’t seen [Albert] for months. (name)

(34) The West is suspecting Iran of building nuclear arms. But negotiations with [Teheran] continue. (metonymy / synecdoche)

(35) [Paul [sings under the shower]_k]_i
   a. Mary finds [that]_i weird. (abstract anaphora)
   b. John does [it]_k, too.
**r-given-sit** (situationally evoked)
Symbolic deixis, uniqueness relative to communicative situation

(36) [I] want [us] to return the car.
(37) [Last week], she told [me] the opposite.
(38) Come [here]!

We do not annotate temporal quantifiers like *always, often, every Wednesday*
Snowden interview: deixis and anaphora

- Abstract anaphor: antecedent of *it* highlighted
- Additional feature + *generic* marks that the expression does not refer to a specific individual but has a class reading.
r-environment

- Gestural deixis, uniqueness of visual object ensured by demonstration
- Occurs only in face-to-face communication

(39) You should take [this way].
(40) [He] kicked me.
**r-bridging** (associative, mediated)

- Discourse-new but dependent on previous context
- Uniqueness within scenario / frame
- An expression with an implicit argument

(41) When they tried to enter the house, the door fell off.
(42) The city is planning a new townhall and [the construction] will start early next year.
(43) Our correspondent in Egypt is reporting that [the opposition] is holding a rally [against the constitutional referendum].

Note that bridging is not *defined* in terms of a part-whole relation!
Snowden interview: bridging anaphora

Bridging anaphors have an antecedent (sometimes silent) that is understood as an implicit argument.

(44) on the inside (of the NSA)
(45) into (your) work

Assign a label to an entire phrase:
r-unused-known

- Unique expression in the global context (first mention)
- Likely to be known by the intended audience

(46) [The Pope] stood [on St. Peter’s Square].
(47) [Space probe Voyager 1] passed [planet Jupiter] [in 1979].
(48) [Igor Stravinsky] died [in New York] and was buried [in Venice].

- Note that the question whether an entity is known by the audience is not a linguistic question but varies over time and for different addressees.
- There are also globally unique entities which are unknown.
Snowden interview: known expression

You were working till last summer for the NSA and during this time.
Complex phrases

- Often, several referring expressions are nested inside each other.
- Each of them receives its own label.

(49) [All activities [on the international airport [in the vicinity]]] came to a halt.

- The same goes for possessive pronouns.

(50) He welcomed them [to his house].
Annotation on available syntactic structures

DIRNDL corpus (Eckart et al. 2012; Björkelund et al. 2014), SALTO tool (Burchardt et al. 2006)
Complex expressions in EXMARaLDA

snowden-interview-answer8-snowden-anno.exb

- Create a label tier for each level of embedding.
r-unused-unknown

- Unique expression in the global context (first mention)
- Not likely to be known by the audience
- Sufficient descriptive material to ensure identifiability and to introduce a unique new entity to the common ground

(51) [The woman Max went out with last night] is an astrophysicist.

(52) [The swimming pool of the new town hall] created discontent among the voters.

(53) I just saw [the creepy reptile of my office colleague].

(54) [Carl, my neighbour,] never gets up before 10 o’clock.
**r-bridging-contained containing inferrable**

- Special type of bridging anaphor: the anchor / antecedent is a syntactic argument of the head noun, i.e. it is embedded in the phrase.

(55) [The opening day of the G20 summit] was a desaster.
(56) We were surprised to even see [the President of Malta].
(57) [The construction of the new townhall] will start early next year.
Distinguishing r-unused-unknown and r-bridging-contained

**Permutation test:** if there is an embedded argument, front it. If the resulting discourse is felicitous, assign the label *r-bridging-contained*. Otherwise, assign *r-unused-unknown*.

(58) a. Markable: [the President of Malta]
   b. Permutation: ✓ I was in Malta and met the President.
      ⇒ *r-bridging-contained*

(59) a. Markable: [the creepy reptile of my office colleague]
   b. Permutation: ??When my office colleague left the room, [the creepy reptile] attacked.
      ⇒ *r-unused-unknown*
r-new

- Non-unique, discourse-new expression

(60) [One stormtrooper] threatened me.

(61) [A military spokesman] confirmed [explosions] and the death [of at least two soldiers].

(62) I’m married [to a computer scientist].
References


