Is intransitive subject the preferred role for introducing new referents?
Evidence from corpus-based typology

Stefan Schnell  
ARC CoEDL / U Melbourne

Nils Norman Schiborr  
University of Bamberg

Geoffrey Haig  
University of Bamberg

01 September 2018
Overview

1. Introduction:
   Argument structure as architecture for referent introduction

2. Language corpora, annotations, and methods

3. Findings:
   P arguments as the predominant entry point for new referents

4. Conclusions and outlook
Once upon a time...

(1) Grimm’s *Fairy tales*, “The wolf and the seven young kids”
   a. *There was once upon a time [an old goat]_S*
   b. *[who]_A had [seven little kids]_P*
   c. and *Ø_A loved [them]_P* ...

(2) *Once upon a time [an old goat]_A had [seven kids]_P* ...
… most lexical mentions occur in absolutive argument positions (S or O [= P]), but are avoided in the ergative (A) slot, which is mostly restricted to reduced forms (pronoun, agreement, zero). Correspondingly, most new mentions occur in S or O [= P], with few occurring in A.

– Du Bois 2017: 29, emph. added
Preferred argument structure

- certain syntactic roles are systematically associated with particular information status
- due to assumed constraints on information processing
  Chafe’s (1987) ‘one new concept at a time constraint’
  Lambrecht’s (1994) ‘separation of role and reference’

cf. ‘pragmatic linking’ (Durie 2003):
S + P provide “a predictable locus for unpredictable work”
(Du Bois 2003a: 47)
Among the things that a speaker may know about the verb come in, for example, is that its S role provides a reliably usable slot for introducing a new human protagonist into a discourse. Likewise, the O [= P] of the verb meet may serve a similar function.

– Du Bois 2003a: 40, emph. added
Cognitive constraints

The way to fulfilment [i.e. to abiding by the constraints on argument structure in discourse processing] is via a simple principle of discourse: Speakers need not say everything in one clause. Facing cognitive constraints that could frustrate their expressive goals, speakers can simply mobilize their planning capacity to organize a series of successive clauses.

– Du Bois 2003b: 76, emph. added
**Sakapultek** (Mayan, Du Bois 1987: 822; 828)

### A role
- (n = 187)  
- New: 3
- Lexical: 6

### S role
- (n = 258)  
- New: 22
- Lexical: 48

### P role
- (n = 170)  
- New: 25
- Lexical: 46
Intransitive introduction

- but, it is **intransitive clauses** that are added for the sake of information flow rather than for their “conceptual content or semantic one-placeness” (Du Bois 1987: 831)
- e.g. predicates like *arrive, appear, come in*
- hence ‘**intransitive introduction, transitive narration**’: speakers prefer the S role for introducing new referents to minimize processing costs
Intransitive introduction

(3) *Pear stories* (Chafe 1980)

a. *a man was picking pears*  
(p. 316, speaker 16)

b. *there was a man, who was picking pears*  
(p. 306, speaker 7)

- both options are **grammatical**,  
- but the second is **preferred**
Research questions

1. Is the **S role**, as claimed, cross-linguistically “specialized” for the **introduction of new referents**, regardless of language type?
Research questions

1. Is the S role, as claimed, cross-linguistically “specialized” for the introduction of new referents, regardless of language type?

2. In what ways can this be meaningfully evaluated from a statistical perspective?
1. Is the **S role**, as claimed, cross-linguistically “specialized” for the **introduction of new referents**, regardless of language type?

2. In what ways can this be meaningfully evaluated from a **statistical perspective**?

3. Are the **S and P roles** comparable in this respect?
1. Is the **S role**, as claimed, cross-linguistically “specialized” for the **introduction of new referents**, regardless of language type?

2. In what ways can this be meaningfully evaluated from a **statistical perspective**?

3. Are the **S and P roles** comparable in this respect?

4. What is the role of **non-core arguments** in managing new information?
The sample

- non-elicited, monologic spoken narratives from the freely accessible **Multi-CAST collection** (Haig & Schnell 2015)

<table>
<thead>
<tr>
<th>corpus</th>
<th>affil.</th>
<th>n(clauses)</th>
<th>citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cypriot Greek</td>
<td>I.-E.</td>
<td>1 071</td>
<td>Vollmer &amp; Hadjidas 2015</td>
</tr>
<tr>
<td>English</td>
<td>I.-E.</td>
<td>1 244</td>
<td>Schiborr 2015</td>
</tr>
<tr>
<td>North. Kurdish</td>
<td>I.-E.</td>
<td>1 389</td>
<td>Haig &amp; Thiele 2015 *</td>
</tr>
<tr>
<td>Sanzhi Dargwa</td>
<td>Nakh-Dagh.</td>
<td>1 702</td>
<td>Forker &amp; Schiborr in prog.</td>
</tr>
<tr>
<td>Teop</td>
<td>Oceanic</td>
<td>1 272</td>
<td>Mosel &amp; Schnell 2015</td>
</tr>
<tr>
<td>Vera’a</td>
<td>Oceanic</td>
<td>2 377</td>
<td>Schnell 2015</td>
</tr>
<tr>
<td><strong>totals</strong></td>
<td></td>
<td><strong>8 971</strong></td>
<td></td>
</tr>
</tbody>
</table>

* with additional contributions by Maria Vollmer
Methodology: annotations

- the corpora have been manually annotated for
  - the form and role of referring expressions, (with GRAID, Haig & Schnell 2014)
  - the identity of each mention of a referent, and (with RefIND, Schiborr et al. 2018)
  - the information status of new referents (with RefLex, Riester & Baumann 2017)
Examples

(4) **Sanzhi Dargwa** [sanzhi_devil_034]

\[
\begin{align*}
\text{xun-ne-b} & \quad \text{suk} & \quad \text{b-ič-ib} & \quad \text{k:urt:a} \\
\text{road-SPR-N} & \quad \emptyset & \quad \text{meet} & \quad \text{N-occur.PFV-PRET} & \quad \text{fox} \\
\text{np:l} & \quad 0.h:s & \quad \text{other} & \quad \text{v:pred} & \quad \text{np.d:p} \\
\text{0002} & \quad & \quad & \quad & \quad \text{0031} \\
\text{new}
\end{align*}
\]

‘On the road (he) met a fox.’
GRAID: grammatical relations

- defined as per Andrews (2007, 1985) as ‘generalizations of semantics prototype roles across encoding properties’
  - $S$ = subjects of intransitive clauses
  - $A$ = function coded like prototypical agents
  - $P$ = function coded like prototypical patients
GRAID: grammatical relations

- additionally, we distinguish
  - **goals**, recipients, and addressees,
  - static **locations**,
  - other **oblique** arguments, and
  - others (possessors, predicates, direct address, etc.)
RefIND: identifying discourse referents

- **discourse referents**
  - are linguistic representations of construed entities,
  - time-stable within the universe of discourse and
  - trackable across states-of-affairs throughout a discourse (Du Bois 1980); but which
  - exclude various instances of nominal expressions,
    - e.g. those under scope of negation,
    - predicates of classificatory clauses,
    - conflated objects, etc.
RefINd: new vs. given

- at the first mention in linear order of a particular referent, that referent is considered **newly introduced**;
- all subsequent mentions of the same referent are assumed to be **given**
RefLex: new, unused, and bridging

bridging
a. referent inferable from frame semantics or
b. a previously mentioned situation, or
c. anchored to an already given referent

unused
a globally known entity,
via encyclopædic or cultural knowledge

(brand) new
referent not otherwise inferable or known
Examples

(5) **Sanzhi Dargwa** [sanzhi_devil_038]

\[
\begin{align*}
\text{k:urt:a-l} & \quad \text{b-ič:-ib} & \quad \text{hel-i-j} & \quad \text{cin-na} & \quad \text{bez} \\
\text{fox-ERG} & \quad \text{N-give.PFV-PRET} & \quad \text{that-OBL-DAT} & \quad \text{REFL.SG-GEN} & \quad \text{hair} \\
\text{np.d:a} & \quad \text{v:pred} & \quad \text{pro.h:g} & \quad \text{ln_refl.d:poss} & \quad \text{np:p} \\
0031 & \quad 0002 & \quad 0031 & \quad 0032 & \quad \text{bridging}
\end{align*}
\]

‘The fox gave him one of its hairs.’
Methodology: automation

- based on the annotations, we can algorithmically determine
  - the **frequency** of each unique referent,
  - the **position** of each mention relative to others (e.g. for newness, lookback distance, etc.), and
  - the **relative proportions** of each group (i.e. corpus | role | information status)
Methodology: procedure

1. The texts are annotated in ELAN,* stored as XML files

2. A custom R script reads these XML files, and transforms them into a table (data and scripts are available in the multicastR package**)

3. Each row of the table represents one ‘grammatical word’, i.e. the smallest GRAID annotation unit

4. The table is filtered for the heads of referring expressions, thereby excluding all non-referring material, so that one data point = one referring expression

* https://tla.mpi.nl/tools/tla-tools/elan/
** https://CRAN.R-project.org/package=multicastR
Given and new: proportions

- tracked referents: 1,273 across 29 texts
- total mentions: 13,677
  - brand new: 531 (4%)
  - bridging: 565 (4%)
  - given: 12,533 (92%)

- ‘tracked’ referent
  = unique referent with at least two mentions
  (excluding 1104 referents mentioned only once)
given a newly introduced referent, what is the probability of it being in a particular role?
A role (n=2895)  

S role (n=4591)  

P role (n=2193)  

% of non-given mentions
what proportion of each role is dedicated to newly introduced referents?
A role
(n = 2895)

S role
(n = 4591)

P role
(n = 2193)

% non-given mentions

C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera’a

C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera’a

C. Greek
English
N. Kurdish
S. Dargwa
Teop
Vera’a

2
5
2
2
2
1

6
8
4
11
3
1

24
28
13
20
10
12
<table>
<thead>
<tr>
<th>Language</th>
<th>Goals (n=715)</th>
<th>Locations (n=314)</th>
<th>Obliques (n=397)</th>
<th>Other (n=2572)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. Greek</td>
<td>31</td>
<td>29</td>
<td>29</td>
<td>8</td>
</tr>
<tr>
<td>English</td>
<td>13</td>
<td>16</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>N. Kurdish</td>
<td>18</td>
<td>25</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>S. Dargwa</td>
<td>21</td>
<td>24</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Teop</td>
<td>14</td>
<td>9</td>
<td>28</td>
<td>8</td>
</tr>
<tr>
<td>Vera’a</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

% non-given mentions

- C. Greek
- English
- N. Kurdish
- S. Dargwa
- Teop
- Vera’a
The prominence of P

- consistently across all corpora,

\[ p(P_{\text{new}}) > p(S_{\text{new}}) > p(A_{\text{new}}) \]

(Fisher’s exact tests for S and P yield \( p < 0.0001 \) for all corpora)

- in general, the P role
tends to have the largest fraction of mentions of any core argument dedicated to new introductions
New vs. bridging

- discourse-new referents that are evoked/inferable (bridging) should be easier to process, hence require less effort to place into predictable loci

- generally, bridging should thus
  - be more freely associated with the constraint role of A,
  - and likely less frequently in P and the processing-aiding position S,
  - than (brand) new referents
what proportion of each role is dedicated to brand new and bridging introductions?
A role (n=2895)

S role (n=4591)

P role (n=2193)

% non-given mentions

C. Greek English N. Kurdish S. Dargwa Teop Vera’a

1 0 0 0 1 1

3 3 2 5 2 0

12 17 5 11 6 8

1 5 1 2 1 1

3 6 2 5 2 1

11 11 8 9 4 5

C. Greek English N. Kurdish S. Dargwa Teop Vera’a

0 5 10 20 30

0 5 10 20 30

0 5 10 20 30

0 5 10 20 30

0 5 10 20 30

new (n=531)

bridging (n=565)
Conclusion: S and new information

- the S role appears “specialized” for new introductions only by the proportion of S among all new introductions
Conclusion: S and new information

- the S role appears “specialized” for new introductions only by the proportion of S among all new introductions

- but,
  S arguments are overall highly frequent, hence offering more potential ‘landing sites’ for new information
Conclusion: S and new information

- The S role appears “specialized” for new introductions only by the proportion of S among all new introductions.

- But, S arguments are overall highly frequent, hence offering more potential ‘landing sites’ for new information.

- From an intra-role perspective, new mentions make up only a small fraction of the S arguments in a text.
Conclusion: P and new information

- instead, the P role (and certain non-core arguments) harbour much larger proportions of new mentions

- for the P role especially, we might hypothesize a genuine cross-linguistic association with new information
Conclusion: P and new information

- this association may in turn be motivated by the association of new mentions with certain semantic roles (rather than by pragmatic linking to a syntactic position)

- also:
  linking of a new referent to an already established one in a transitive construction
Conclusion: non-core functions

- **non-core functions** generally harbour high proportions of new mentions

- **goals/recipients/addressees** tend to bear fewer new mentions (especially when human)

- than **locations and other oblique arguments** (in particular when non-human)

- **however,**
  large inter-corpus variability within these roles (possibly more content-sensitive?)
Conclusions: new vs. bridging

- no clear effect

- but again,
  
  it is P, but not S, that is most inclined to host brand-new mentions!
Corpus-based typology

- challenges:
  - size of data sets and representativeness
  - comparability of data sets
  - inter-annotator differences
  - open availability of data and methods
  - also: annotation workload
Going forward

- **with our established methods**, we can readily determine **anaphoric relations** (e.g. lookback distances, role continuity) as well as **related anaphoric forms** and their **syntactic functions**, and more
Going forward

- furthermore, a **notion of ‘topic’** can be secondarily derived,
- e.g. as **expressions** that
  - a. occur above a certain frequency threshold;
  - b. occur in S or A role in strings of consecutive clauses;
  - c. are most frequently realized via reduced forms (pronouns, zero);
  
  etc.
all data will (in the near future) be freely accessible online at

https://lac2.uni-koeln.de/multicast/

— normally at —

https://lac.uni-koeln.de/multicast/


References (2/3)


References (3/3)


