A parallel corpus approach to sudah

Bruno Olsson & David Moeljadi

ISMIL 18, Procida, June 2014

Introduction

- *sudah* 'already' is the most frequently occurring aspectual marker in Indonesian
- Probably also the most researched aspect marker
- Markers meaning 'already' seem to play a major role in the aspect systems of (all?) languages of Southeast Asia
  - Cross-linguistic studies
- Basic meaning 'already'
  - *sudah* etc. typically treated as a kind of perfect
  - Shares the feature of "Current Relevance" with European perfects
- Our goal: Investigate the contexts that trigger the use of *sudah* and compare this to the use of the perfect
- Explore parallel subtitles as a data source

A parallel corpus of English–Indonesian movie subtitles

- Parallel texts have become an important data source for typology
- Parallel movie subtitles have not been used
- Some advantages:
  - Easily accessible in machine readable format, only minimal preprocessing required
  - Closer to 'natural speech' than the Bible or newspaper text
  - Most of the text is direct speech
- Disadvantages:
  - Copyright issues
  - More features of written language than we expected…
- Preparation of the corpus:
  - Subtitles collected from a popular Internet source
  - Mark-up (time stamps, font information) and metadata removed
  - Line-to-line alignment using Hunalign
  - Translation and original were lumped together
- Total ca. 125000 words

Summary of the corpus

Table : Partial summary of the corpus.

Title	Lang.	*n* lines	*sudah*/ 1000lines	PERF/ 1000lines
Harry Potter	ENG	1230	31	55
Titanic	ENG	1990	32	30
Avatar	ENG	1246	36	16
Man of Steel	ENG	1272	24	48
The Last Stand	ENG	1000	26	22
Laskar Pelangi	IND	823	60	43
Habibie & Ainun	IND	1204	56	39
Batas	IND	882	40	52
Serbuan Maut	IND	469	60	41
Total		**10116**		

Semantics of the perfect

- Functions of the perfect:
  - **Perfect of Result**:
    I have locked the door
  - **Perfect of Persistent Situation**:
    He has been writing for an hour
  - **Experiential**:
    John has been to Paris
  - **Hot News**:
    The king has just died
- What connects these functions is the Current Relevance

Semantics of *sudah* and *already*

- *Phasal* markers, making reference to a positive phase, and presupposing an earlier negative state, separated by a transition ('change of state').
  - Strange with predicates that are not preceded by a negative state, as in *The eggs are already raw*
- Ebert (2001) refers to this as a 'new situation' and proposed the term NEWSIT to cover 'already'-markers.
- Not clear how basic the 'earlier than expected' component is
- That *sudah* has "current relevance" is clear from its phasal meaning. But *why* speakers chose to add *sudah* to a sentence (stative readings are often available without *sudah*).

Our findings

- *sudah* and the perfect are almost equally frequent in the data: 383 vs. 377 occurrences respectively
  - However, the overlap of the occurrences is small: *sudah* and the perfect occur in corresponding lines 114 times, jaccard distance 0.18 (with 1 being complete overlap)
- **No connection between resultativity and use of *sudah*.** The 235 instances of Perfects of Result make up the majority (62%) of perfects in the English corpus. In fact, 85 of these are translated using *sudah* (20% of all instances of *sudah*) but never because of resultativity *per se*.

(1) HARRYPOTTER475

The problem is, I can't remember what I've forgotten.	Masalahnya, aku tidak ingat apa yang aku **(Ø)** lupakan.

Findings (cont.)

- Example of *sudah* corresponing to a Perfect of Result:
  (2) has a resultative (*have seen*) translated with *sudah*, not to convey the result state but rather to cancel A's presupposition that B has not yet seen them:

(2) LASTSTAND6269

A.	— Oh, really? You think so?	— Oh ya? Menurutmu begitu?
B.	— I know so. I've seen them.	— Memang begitu, aku **sudah** melihatnya.

Findings (cont.)

- We treated only instances of perfect plus the word *just* as Hot News. This construction corresponds to Indonesian *baru saja*, which is incompatible with *sudah*.

(3) LASKAR7059

I've just seen the prettiest fingernails in the world!

Aku baru saja (Ø) lihat kuku paling cantik sedunia!

Findings (cont.)

- **The most frequent use of *sudah* is in contexts involving a 'local' expectation**: 123 instances of *sudah* (32%) were classified as such. A typical context is questions about whether something has been done yet:

(4) MANOFSTEEL5588

Ship, have you managed to quarantine this invasive intelligence?

Pesawat, **sudah**kah kau mengkarantina kecerdasan asing ini?

(5) MANOFSTEEL5557

You should have visual contact now.

Seharusnya kau **sudah** bisa melihatnya sekarang.

Findings (cont.)

- Another frequent use: **natural developments**, i.e. when a state is the natural outcome of a process such as the change between night and day, or life and death: 45 occurrences (12%).

(6) HABIBIE8240

Sir, it's morning. Okay.	**Udah** pagi. Pak. Ah. ya. oke.

(7) HARRYPOTTER787

How can it? Both my parents are dead.	Bagaimana mungkin? Kedua orang tuaku **sudah** meninggal.

- These are related to the "local expectation" cases above; however, we treat them as a separate category since these expectations seem to be on a higher level than the immediate speech context.

Findings (cont.)

- Convergece between *sudah* and PERFs in what we call **cumulative contexts**: sentences referring to a stage that has been reached by accumulation. A clear example is Perfect of Persistent Situation, which typically corresponds to sentences with *sudah*:

(8) LASTSTAND5194

Been here for 33 years, doctor.	Aku **sudah** di sini selama 33 tahun, Doktor.

(9) TITANIC1740

I've been on my own since I was 15.	Aku **sudah** mandiri sejak berumur 15 tahun.

Findings (cont.)

- Cumulatives also refer to how many times something has occurred, or what amount has been reached:

(10) HARRYPOTTER744

We've looked a hundred times!	Kita **sudah** cari ratusan kali!

(11) TITANIC2389

Boiler Room 6 is flooded 8 feet above the plate…	Ruang Pemanas 6 **sudah** banjir 8 kaki di atas plat…

- It is tempting to see the connection between expectations, natural developments and cumulatives as involving unidirectional changes
- These contexts are also compatible with "event focusing" readings if *sudah* is not present, giving dynamic interpretations

Findings (cont.)

- We also identified a number of other, less frequent, uses:
  - The use of *sudah* for events that occur **earlier than expected**, 13 instances (3%):

(12) HABIBIE7778

You just started your leave, and you already found your love.

Kamu ini baru aja cuti **sudah** dapat jodoh.

  - A handful of what we call **non-permanent state**:

(13) HARRYPOTTER89

Daddy's gone mad, hasn't he?

Ayah **sudah** gila, ya?

Findings (cont.)

- Finally, there is number of perplexing uses that we have failed to connect to other uses or find any independent explanations for. One example is what we call the **terima kasih sentences**:

(14) BATAS9566

Thank you for taking me up to here.

Terima kasih **sudah** mengantar aku sampai di sini.

Conclusions

- Using a parallel movie subtitle corpus is a relatively easy way to compare grammatical categories across languages.
- We found little overlap between the perfect and *sudah*; the observed convergence is due to different reasons for each category, making it difficult to claim that they are both instances of the same cross-linguistic category.
- The really interesting part remains: figuring out how the use of *sudah* differs from the use of similar markers in other languages of Southeast Asia.