ISMIL 18, Procida, June 2014

A parallel corpus approach to *sudah*

Bruno Olsson & David Moeljadi
olssonbruno@gmail.com, davidmoeljadi@gmail.com

1 Introduction

- *sudah* 'already' is clearly the most frequently occurring of the aspectual markers in Indonesian and other Malay varieties (*udah*, *dah*, *su*. . .).
- Probably also the most researched aspect marker (e.g. Gonda 1954, Minde and Tjia 2002, Grangé 2010, Soh 2012).
- Markers meaning 'already' seem to play a major role in the aspect systems of (all?) languages of Southeast Asia: Thai (Jenny 2001), Lao (Enfield 2007), Burmese (Okell 1969), Vietnamese (Do-Hurinville 2004), Javanese (Vander Klok and Matthewson 2014), etc.; also similar markers in Himalayan languages (Ebert 2001), Chinese (Li et al. 1982) and various Oceanic languages (e.g. Mwotlap, François 2003).
  - Attempts at cross-linguistic studies: Olsson 2013, Dahl and Wälchli 2013
- Researchers seem to agree that these markers mean 'already'. But this does not account for (a) the large functional load and (b) the cross-linguistic differences in the use of the markers.
  - Solution is typically to treat *sudah* etc. as a kind of perfect
  - Shares the feature of "Current Relevance" with European perfects, although what counts as 'relevance' seems to differ (cf. Li et al. 1982 for Mandarin Chinese *le*)
- Our goal: Investigate the contexts that trigger the use of *sudah* and compare this to the use of the perfect
- We also hope that the method can be used for future studies including subtitles in more languages, in order to investigate the differences between Indonesian and other languages with similar aspectual markers. Subtitles are relatively easy to find for Thai, Vietnamese and Mandarin, which is promising for future research.

2 A parallel corpus of English–Indonesian movie subtitles

- Parallel texts have become an important data source for typological investigations, most importantly translations of the New Testament (Cysouw and Wälchli 2007).
- Parallel movie subtitles have not been used for typological purposes (as far as we know). Several 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: corpus cannot be shared
    - The texts turned out to have more features of written language than we had expected, e.g. 76 instances of *telah* (still better than the NT).

- Preparation of the corpus:
    - Subtitles were collected from a popular Internet source.[1] Titles were chosen opportunistically: a few blockbuster movies in each language, and corresponding translations. A stricter approach should sample according to genre, year of production, etc.
    - The files were preprocessed: mark-up (time stamps, font information) and metadata removed.
    - Line-to-line alignment was performed using Hunalign (Varga et al. 2005).

- For our purposes we make no difference between translation and original, and simply lump them together. A stricter approach would treat them as non-equivalent:
    - Lower frequency of aspect markers in Indonesian 'translationese': *sudah* is much more frequent in the originals (mean = 54 occurrences per 1000 words) than in the translations (mean = 30 occurrences per 1000 words; significant at $p < 0.01$).
    - Difference between frequency of PERFs in original and translations not significant (large variation).

The contents of the corpus are summarized in Table 1.

3 Semantics of the perfect

- Research on the perfect typically lists different types/functions of the perfect (e.g. Comrie 1976: 56-61):
    - **Perfect of Result** is the historical source (combined with the possessive *have*-construction) of the European perfects, and remains its prototypical use: *I have locked the door*.
    - **Perfect of Persistent Situation** or 'Universal Perfect': *He has been writing for an hour*.
    - **Experiential** is used for an event that occured at least one in the past (cf. repeatability constraint), typically used when the subject gained some sort of knowledge from the event ('experience'): *John has been to Paris*. Frequently in non-affirmative contexts (questions, negated).
    - **Hot News** is more dubious than the others, as it is not clear what distinguishes *The king has just died* from the Perfect of Result except the minimal time frame.

[1] http://www.subscene.com

Table 1: Summary of the corpus.

Title	Lang.	n words ENG/IND	n lines	n *sudah* (/*udah*)	n PERF	*sudah*/ 1000lines	PERF/ 1000lines
Harry Potter and the Philosopher's Stone	ENG	13450/7188	1230	38	68	31	55
Titanic	ENG	13168/10590	1990	64	59	32	30
Avatar	ENG	8296/6496	1246	45	20	36	16
Man of Steel	ENG	8334/6463	1272	30	61	24	48
The Last Stand	ENG	6819/4994	1000	26	22	26	22
Laskar Pelangi	IND	6244/5387	823	49	35	60	43
Habibie & Ainun	IND	6927/5524	1204	60/8	47	56	39
Batas	IND	5445/4063	882	23/12	46	40	52
Serbuan Maut	IND	2869/2695	469	13/15	19	60	41
Total		**71552/53400**	**10116**	**348/35**	**377**		

- What connects these uses is the Current Relevance, although this term has to be given a rather vague interpretation to cover all cases.

4 Semantics of *sudah* and *already*

- Both markers can be characterized as *phasal*, making reference to a positive phase, and presupposing an earlier negative state, separated by a transition ('change of state') (see e.g. Löbner 1989, Soh 2012). Strange with predicates that are not preceded by a negative state, as in *The eggs are already raw*, which is harder to fit into a suitable context than *The eggs are already cooked*.
- 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. Vander Klok and Matthewson (2014) seem to treat the "earliness implicature" as a fundamental part of Javanese *wis*, but van der Auwera (1998) shows that languages differ as to whether 'already'-words give rise to such meanings. Probably safer to treat 'earlier than expected' as one of several contexts in which such words can be used.
- That *sudah* has "current relevance" is clear from its phasal meaning, as it asserts a state while backgrounding any event that led to this state. An important question however, is *why* speakers chose to add *sudah* to a sentence (stative readings are often available without *sudah*).

5 Our findings

- It is striking that *sudah* and the perfect are almost equally frequent in the data (383 vs. 377 occurrences respectively), suggesting a similar functional load. However, the overlap of the occurrences is small:

sudah and the perfect occur in corresponding lines 114 times, giving a jaccard distance of 0.18 (with 1 being complete overlap). If we assume that items with the same function/meaning should have the same distribution across translations, there are clearly differences between the two.

- **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 we would argue that this is for other reasons than 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.

Example (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.

- Most co-occurrences of resultatives and *sudah* are either from such **presupposition cancelling contexts** (these make up a total of 19% of all instances of *sudah*), or contexts were the event producing the result state is supposed to take place within the relevant time frame (see below).
- Since the delimitation of the Hot News perfect was not clear from the literature, we decided to treat only instances of perfect plus the word *just* as Hot News. This construction corresponds to Indonesian *baru saja*, which is incompatible with *sudah*, presumably because *baru saja* is incompatible with stative predicates in general.

(3) LASKAR7059

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

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

- **The most frequent use of *sudah* is in contexts involving a 'local' expectation**, e.g. that the participant(s) expected the event to occur within the time frame. 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.

- Another frequent use, clearly related to the preceding, involves what we call **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:

(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 since 'being morning' or 'being dead' are also expected outcomes; however, we treat them as a separate category since these expectations seem to be on a higher level than the immediate speech context. Predicates such as *meninggal* or *mati* occur with *sudah* when used to inform about the status of people, but *mati* would not be used with e.g. electrical appliances since these do not undergo the same natural developments (*AC-nya mati* 'The air-con is off'), unless there is a local expectation that the air-con should get turned off.

- One of the most interesting findings, in our view, is the degree to which *sudah* and PERFs converge in what we call **cumulative contexts**. By this we mean a sentences referring to a stage that has been reached by accumulation. Typically this involves an object or adverbial modifies by a numeral. 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.

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, i.e. processes or changes that occurs in a given order (the occurrence of an expected event, the accumulation of time, etc.).

- These contexts are also compatible with "event focusing" readings if *sudah* is not present, giving dynamic interpretations. It could be that *sudah* felt to be required to emphasize the stativity in contexts that are likely to receive both static and dynamic readings.

- We also identified a number of other, less frequent, uses:

  - The use of *sudah* for events that occur **earlier than expected** turned out to be rare. Counting generously, we only found 13 instances (3%). Here is a relatively clear example:

(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** (for lack of a better term) where we interpret *sudah* as simply conveying that the state is "new", and did not hold before:

(13) HARRYPOTTER89
Daddy's gone mad, hasn't he?
Ayah **sudah** gila, ya?

- 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 **terimah kasih sentences**:

(14) BATAS9566
Thank you for taking me up to here.
Terima kasih **sudah** mengantar aku sampai di sini.

6 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.

References

Comrie, Bernard. 1976. *Aspect*. Cambridge: Cambridge University Press.

Cysouw, Michael and Bernhard Wälchli. 2007. “Parallel texts: Using translational equivalents in linguistic typology”. In: *Sprachtypologie und Universalienforschung* 60.2, pp. 95–99.

Dahl, Östen and Bernhard Wälchli. 2013. “Disentangling the variability of the perfect gram type”. Presented at the ALT 10, Leipzig.

Do-Hurinville, Danh Thành. 2004. “Temps et aspect en vietnamien: Étude comparative avec le français”. PhD thesis.

Ebert, Karen H. 2001. “Tense-aspect flip-flop and a somewhat elusive gram type”. In: *Aktionsart and aspectotemporality in non-European languages*. Ed. by Karen H. Ebert and Fernando Zúñiga. Zürich: ASAS-Verlag, pp. 141–158.

Enfield, Nick J. 2007. *A Grammar of Lao*. Berlin: Mouton de Gruyter.

François, Alex. 2003. *La sémantique du prédicat en mwotlap (Vanuatu)*. Leuven, Paris: Peeters.

Gonda, Jan. 1954. “Tense in Indonesian Languages”. In: *Bijdragen tot de Taal-, Land- en Volkenkunde* 110.3, pp. 240–262.

Grangé, Philippe. 2010. “Aspect and modality in Indonesian: The case of *sudah*, *telah*, *pernah*, *sempat*”. In: *Wacana* 12.2, pp. 143–168.

Jenny, Mathias. 2001. “The aspect system of Thai”. In: *Aktionsart and aspectotemporality in non-European languages*. Ed. by Karen H. Ebert and Fernando Zúñiga. Zürich: ASAS-Verlag, pp. 97–140.

Li, Charles N. et al. 1982. “The Discourse Motivation for the Perfect Aspect: The Mandarin Particle le”. In: *Tense-Aspect: Between semantics & pragmatics*. Ed. by Paul J. Hopper. Benjamins, pp. 19–44.

Löbner, Sebastian. 1989. “German *schon – erst – noch*: An integrated analysis”. In: *Linguistics and Philosophy* 12.2, pp. 167–212.

Minde, Don van and J. Tjia. 2002. “Between Perfect and Perfective. The meaning and function of Ambonese Malay su and suda.” In: *Bijdragen tot de Taal-, Land- en Volkenkunde* 158.2, pp. 283–303.

Okell, John. 1969. *A Reference Grammar of Colloquial Burmese (two volumes)*. London: Oxford University Press.

Olsson, Bruno. 2013. “Iamitives: Perfects in Southeast Asia and beyond”. Masters thesis. Stockholm University.

Soh, Hooi Ling. 2012. “Towards a semantic analysis of the aspectual marker *dah* in Colloquial Malay”. Presented at the 16th International Symposium on Malay/Indonesian Linguistics, Kelaniya (Sri Lanka).

van der Auwera, Johan. 1998. “Phasal adverbials in the languages of Europe”. In: *Adverbial constructions in the languages of Europe*. Ed. by Johan van der Auwera and Dnall P. Baoill. Berlin: Mouton de Gruyter.

Vander Klok, Jozina and Lisa Matthewson. 2014. “Distinguishing *already* from perfect aspect: A case study on Javanese *wis*”. Presentation given at AFLA 21, University of Hawaii.

Varga, D. et al. 2005. “Parallel corpora for medium density languages”. In: *Proceedings of the RANLP 2005*, pp. 590–596.