Corpus-based approach meets LFG: Puzzling voice alternation in Indonesian

Gede Primahadi Wijaya Rajeg [1], I Made Rajeg [1], & I Wayan Arka [1, 2]
Universitas Udayana, Indonesia [1] & Australian National University [2]

This paper discusses a novel approach to the study of voice in Indonesian. Our goal is to provide fresh, corpus-based evidence that voice alternations are not always meaning-preserving argument structure alternations (Kroeger, 2005, p. 271). Against the status-quo in linguistic theorising about voice alternation, we argue that voice is a lexical-constructional phenomenon with a particular voice (type) carrying its own constructional semantic traits and idiosyncrasies (cf. Booij, 2010), often susceptible to grammaticalisation. We examine verbs with different transitive (applicative/causative) suffixes, *-kan* and *-i*, building on earlier studies (e.g., Arka et al., 2009, among others). We demonstrate that the lexical-constructional property of Indonesian voice can be naturally handled by the machinery in Lexical Functional Grammar (LFG).

We focus on verbs derived from the root *kena* 'hit; get into contact with', which is associated with negative affectedness and past event/completive aspect. We start with the key puzzling examples shown in (1)-(2): the *-kan* form **mengenakan* (active) is never used in the same sense expressed by *mengenai* in (1) whereas the *-kan* form in *dikenakan* (passive) can convey similar sense expressed by *dikenai* as in (2). That is, the contrast of *–i* and *–kan* in AV as in (1) suddenly disappears in the PASS construction.

(1) *Tak ayal lagi air kotor itu* ***meng-(k)ena-i/*meng-(k)ena-kan*** ***baju*** *Dimas*.
NEG slow again water dirty DEM AV-hit-APPL/AV-hit-CAUS shirt NAME
'Soon enough, that dirty water *hits* Dimas' *shirt*.' (ind_mixed_2012_1M-sentences.txt:774789)

(2) *Sedangkan motor kedua akan* ***di-kena-i*** *pajak sebesar 2 persen.*
meanwhile motor second FUT PASS-hit-APPL tax as.large two percent
'Meanwhile, the second motorbike will be *subject to/charged with* 2% tax.' (ind_mixed_2012_1M-sentences.txt:296558)

The case of possible alternation between *–i* and *–kan* in (2) appears to exemplify what Sneddon et al (2010, p. 101) call the blurry semantic distinction in common usage between *-kan* and *-i* forms of the same root. Sneddon et al. gloss the two AV verbs (*mengenai* and *mengenakan*) as 'subject to' but without providing any examples. Such decontextualised characterisation of these AV verbs contradicts our native speaker intuition. To test our intuition regarding the use of these two AV verbs, we apply a quantitative corpus-linguistic method called *Collostructional Analysis* (CollAna; see below) (Stefanowitsch, 2013). We predict that the construction when both *kenakan* and *kenai* may be synonymous is in the passive *di-* as in (2) rather than in the AV forms (*mengenai/mengenakan*). To test this, we analysed the usage sentences of the passive *dikenai* and *dikenakan* and coded for their senses based on their co-occurring contexts.

The data for this study comes from one corpus file of the *Indonesian Leipzig Corpora Collection* (Quasthoff & Goldhahn, 2013), namely "ind_mixed_2012_1M-sentences.txt". This file is mostly derived from Indonesian online news website (Quasthoff & Goldhahn, 2013, p. 26) and amounts to 15,052,159 million word-tokens. CollAna is a cover term for a family of quantitative methods designed to analyse the interaction between (abstract) grammatical constructions and lexical items that typically occur in (one of) the slot(s) of the constructions (e.g. verbs that occur significantly more often than chance in the ditransitive construction) (Stefanowitsch, 2013). The classes of lexical items typically occurring in the (given slot of the) construction are used to characterise the constructional meaning of the construction. We expand CollAna to test Sneddon et al's claim that *mengenakan* and *mengenai* exemplify the blurry semantics of *-kan*/*-i* verb-pairs in common usage. As a first attempt, we looked at significantly attracted collocates that immediately occur to the right of these verbs (i.e. R1 collocates), approximating the fillers for the verbs' direct-object slots; hence, the pattern [*mengenai/mengenakan* + R1 collocate]. CollAna was performed in R (R Core Team, 2019) with *collogetr* package (Rajeg, 2019).

CollAna reveals that the two verbs attract different classes of R1 collocates, suggesting that their common usage patterns are distinct, requiring to their semantic differentiation. For *mengenai*, its top-20 strongly attracted R1 collocates are mostly abstract nouns. They predominantly refer to MATTER-related nouns (e.g. *hal* 'matter', *hal-hal* 'matters', *rencana* 'plan', *masalah* 'problem', *soal* 'matter') (cf. (3)), but also include WH-words (i.e. *apa* 'what', *bagaimana* 'how', *siapa* 'who') (cf. (4)), suggesting the presence of subordinate, complement clause. Inspecting a sample of sentences for these collocates indicates that *mengenai* has been grammaticalised into an oblique-like marker; it is analysed in this paper as a (verbal) preposition, categorically a P, meaning 'concerning; regarding to; about', marking the Topic or Theme role:

(3) *Dalam bukunya, Darwin tak mampu membahas sepenuhnya* ***mengenai hal*** *ini.*
inside book.3SG.POSS NAME NEG be.able discuss fully concerning matter DEM
'Within his book, Darwin was unable to fully discuss *concerning* this *matter*.' (ind_mixed_2012_1M-sentences.txt:418366)

(4) *Ia tidak ingin teman-temannya tahu* ***mengenai siapa*** *'kakaknya' itu*
3SG NEG want friend~PL know concerning who older.sibling DEM
'(S)he does not want h(is/er) friends to know *regarding who* h(is/er) older sibling is (…)' (ind_mixed_2012_1M-sentences.txt:212649)

Evidence for the grammaticalisation of *mengenai* 'concern' as a preposition in its semantic function marking the (concerned) Topic role in (3)-(4) comes from the fact that, like other prepositions, it has a fixed stucture of P+O. It also lacks a paradigmatic voice opposition (i.e. the passive *dikenai* can never have this grammaticalised sense). It should be noted that it is *mengenai* bearing the locative suffix *-i* (cf. Arka et al., 2009), not *mengenakan*, that has been grammaticalised into a 'concern' marker. This grammaticalisation follows a similar path for CONCERN developed out of locative markers in other languages (Heine & Kuteva, 2002, pp. 201–202). Note that the lexical, non-prepositional meaning of *mengenai* 'to hit; come into contact with' (see (1)) is very rare in the corpus (i.e. 4.03% (288 tokens) out of 7,148 occurrences of *mengenai*).

To further emphasise our goal that voice alternations are not always meaning preserving, we analysed the distribution of the lexical meanings between the passive *dikenai* and AV *mengenai*. For *dikenai*, 89.21% (n=124 tokens) out of 139 cases indicate that its syntactic subject is a party imposed to certain rules (e.g., sanction, tax, fee, punishment) (as in (2)). The lexical meaning of AV *mengenai* never conveys this abstract 'imposing/subject to' sense but only the 'physical touching/hitting' sense as in (1), which is conveyed only in 5.04% (n=7) of the cases of *dikenai* (illustrated in (5)). The remaining tokens for *dikenai* convey the 'being affected (of disease)' sense (n=8). The 'imposing' sense of *dikenai* as in (2) is significantly more frequent than its other uses ($x_2 = 195.29$, $df = 2$, $p_{\text{chi-square}} < 0.001$):

(5) (…) *beberapa orang yang* ***di-kena-i*** *anak panah itu terkapar mati* (….)
several person REL PASS-hit-APPL child arrow DEM PASS.sprawled dead
'Several people who got *hit* by those arrows were sprawled dead' (ind_mixed_2012_1M-sentences.txt:81198)

These distinct quantitative distributions of meanings in active and passive form for *kenai* indicate that passive alternation is a lexical-constructional phenomenon in the sense that the information from the (argument) NPs that the *–i/-kan* verbs co-occur with contributes to the construction of the meanings. The passive *di-* form with the same root exhibits quantitative tendency for conveying a semantic profile quite distinct from the active form. To illustrate it further, our corpus-based study reveals that the AV *mengenakan* is distinct from its counterpart *mengenai* as the former is strongly associated with R1 collocates that all refer to CLOTHING or body-related ACCESSORIES. In this case, *mengenakan* predominantly convey lexical meaning of 'to wear (clothes or accessories)'. Then, manual inspection of all 446 tokens of the passive *dikenakan* reveals that the most frequent meaning (i.e. 56.95% [n=254]) is the 'imposing/subject to' sense (similar profile to *dikenai*), which is significantly more frequent than the passive for 'to wear' (38.12% [n=170]) and the other uses (4.93% [n=22]) ($x_2 = 185.61$, $df = 2$, $p_{\text{chi-square}} < 0.001$).

The LFG analysis to capture these distinct predominant usage patterns of the AV/PASS verbs with *–i/-kan* essentially consists of two parts, (i) lexical entry specification and (ii) interaction and competition (particularly blocking) of different information at the level of morphological/phrasal construction. Both are discussed simultaneously. We assume a traditional morpheme-based analysis of Indonesian morphology, where the affixes including the voice and the transitivisers *–i/-kan* have their entries; see Arka et al (2009) for details. The grammaticalised *mengenai* is a P, having an entry like (6). It says that whole form *mengenai* carries a very specific meaning 'concern'. Like any other entries of preposition, it only has an OBJ in its entry and does not allow a passive alternation. Morphologically, its transparent formal form suggests that the meaning is over time bleached off (i.e. grammaticalised) of the verbal elements of transitive AV verbs (i.e. *meN-* + *kena* + *-i*). *Mengenai* 'concern' competes with *kena*-based forms, which compositionally have more general compositional meanings potentially deriving such a meaning but under the Paninian (or Elsewhere) Principle, the more specific ('concern') meaning wins out, blocking any other morphological structure for the same meaning. This could explain the fact that *mengenakan* cannot be used to express 'concern'. Furthermore, grammaticalisation of *mengenai* into a function word of connective also accounts for the high token-frequency of *mengenai* in this function in the corpus (i.e. 95.91% [n=6,856] of 7,148 tokens).

(6) *mengenai* P (↑PRED) = 'concern<(↑OBJ)>'

The puzzling voice alternation facts in (1)-(2) can be accounted for in the following way. It boils down to how the core meaning of 'affectedness' associated with the root *kena* 'get (negatively[1]) affected/hit' interact with the voice marker (AV *meN-* and PASS *di-*) and applicative/causative *–i/-kan* in the larger construction within/outside words. Space prevents us from spelling out the details but the central point is that physical affectedness (i.e. involving physical contact as exemplified by (1)) is specific to the causative/applicative *–i*; not to the *–kan* counterpart. This is not surprising due to the inherent locative/goal meaning of *–i*; so the entry of the suffix *–i* carries the conceptual element (7) in its LCS (Lexical Conceptual Structure) (cf. Arka et al., 2009). Adding the complexity, the causative/applicative *–kan* also carries an AFFECT element (as it is a transitiviser) but it is about affectedness more generally, for which it overlaps with *–i* when 'locative affectedness' is metaphorical (i.e. abstract) to include examples like (2). This is the 'blurring meaning' characterised as 'subject to' by Sneddon et al (2010, p. 101). Note that this overlap is typically only possible (and attested in our corpus) in the passive construction (*dikenai*/*dikenakan*), not in the active form (*mengenakan/*mengenai*). We analyse it as the effect of blocking of *mengenai*, which in its AV form is specifically associated with the (grammaticalised) verbal-preposition marking 'concern' and the physical affectedness of (7). In the full paper we explore how to precisely capture the effect of 'concrete' and 'abstract' affectedness in LFG through the constraint specification of the nominal type feature ([+/-concrete]), possibly annotated in the (morphological) construction/structure as well as verb/noun entries.

(7) A AFFECT U_i ({TO|FROM}) BE.AT($[LOC]_i$)

Finally, the fact that only the verb *kenakan*, not *kenai*, which can mean '(to) wear' as revealed by CollAna can be straightforwardly captured in LFG. This is an instance of morphological construction (Booij, 2010), where such a meaning is paired with the two morphemes (i.e. the root *kena* suffixed with *-kan*) as a unit in an entry as in (8). Given its meaning, not surprising that it imposes a collocational restriction with its OBJ associated with clothing/accessories. Questions remain, however, for how LFG would capture evidence of quantitative tendency that *kenakan* in 'to wear' sense (in contrast to the 'subject to' sense (2)) is significantly less preferred in PASS *di-* construction, but is more frequent in the AV construction.

(8) *kenakan* V (↑PRED) = 'wear<(↑SUBJ)(↑OBJ)>

In the full paper we discuss the implications of our corpus-based findings in LFG and beyond. The issues include formal LFG representations to capture statistical, usage preferences of different voice types to convey certain meanings, and related questions regarding how deeply entrenched such statistical tendencies in the constructional representations of the verbs are in native speakers' minds. We contextualise the discussion within the emerging trend in combining corpus-based and experimental methods in search of converging and/or diverging evidence of the different usage patterns of morphologically related words.

REFERENCES

Arka, I. W., Dalrymple, M., Mistica, M., Mofu, S., Andrews, A. D., & Simpson, J. (2009). A linguistic and computational morphosyntactic analysis for the applicative *-i* in Indonesian. In M. Butt & T. H. King (Eds.), *Proceedings of the LFG09 Conference*. CSLI Publications.

Booij, G. (2010). *Construction Morphology*. Oxford University Press.

Heine, B., & Kuteva, T. (2002). *World Lexicon of grammaticalization*. Cambridge University Press.

Kroeger, P. R. (2005). *Analyzing Grammar: An Introduction*. Cambridge University Press.

Quasthoff, U., & Goldhahn, D. (2013). *Indonesian corpora* (No. 7; Technical Report Series on Corpus Building). Abteilung Automatische Sprachverarbeitung, Institut für Informatik, Universität Leipzig. http://asvdoku.informatik.uni-leipzig.de/corpora/data/uploads/corpus-building-vol7-ind.pdf

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rajeg, G. P. W. (2019). *collogetr: Collocates retriever and Collocation association measure* (Version 1.1.3) [R]. https://doi.org/10.26180/5b7b9c5e32779

Sneddon, J. N., Adelaar, A., Djenar, D. N., & Ewing, M. C. (2010). *Indonesian reference grammar* (2nd ed.). Allen & Unwin.

Stefanowitsch, A. (2013). Collostructional analysis. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford handbook of Construction Grammar* (pp. 290–306). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195396683.013.0016

[1] The ten most strongly attracted R1 collocates for *kena* identified via CollAna is *pajak* 'tax', *batunya* 'the stone' (parts of idiom *kena batunya* 'get into trouble'), *tipu* 'deceive', *marah* 'angry/anger', *racun* 'poison', *getahnya* 'the resin', *hukuman* 'punishment', *imbasnya* 'the impact/effect', *penyakit* 'disease', *semprot* 'spray' (which can have a metaphoric meaning of 'getting a scolding')