K-SNACS: Annotating Korean Adposition Semantics

Jena D. Hwang
Allen Institute for AI
jenah@allenai.org

Hanwool Choe
Georgetown University
hc563@georgetown.edu

Na-Rae Han
University of Pittsburgh
naraehan@pitt.edu

Nathan Schneider
Georgetown University
nathan.schneider@georgetown.edu

Abstract

While many languages use *adpositions* to encode semantic relationships between content words in a sentence (e.g., agentivity or temporality), the details of how adpositions work vary widely across languages with respect to both form and meaning. In this paper, we empirically adapt the SNACS framework (Schneider et al., 2018) to Korean, a language that is typologically distant from English—the language SNACS was originally designed upon. We apply the SNACS framework to annotate the highly popular novella *The Little Prince* with semantic supersense labels over all Korean postpositions. Thus, we introduce the first broad-coverage corpus annotated with Korean postposition semantics and provide a detailed analysis of the corpus with an apples-to-apples comparison between Korean and English annotations.

1 Introduction

Korean has a grammaticalized category of **postpositions** that includes highly polysemous morphemes that mediate semantic relationships between content words. On their own, they represent humble grammatical markers on nominals, but they play a whale of a role in piecing together the meaning of a sentence. Much like English **prepositions** (or collectively **adpositions**), the semantic relations that they encode range from thematic relationships like agentivity and instrumentality to relative circumstantial information like time, location, or purpose.

In this work we develop a Korean adaptation of an existing annotation schema, SNACS (§2.3), which is specifically geared towards adpositional semantics. The expanded schema details 54 semantic and pragmatic categories called **supersenses** that resolve major ambiguities and generalize across adpositional types. Although the SNACS framework was built based on English preposition senses, the authors claim that the semantically coarse-grained and lexically-agnostic characteristics of the supersenses are well-suited to their adoption for other languages (Hwang et al., 2017). The schema has been so far applied successfully to Mandarin Chinese (Peng et al., 2020). We now apply it to Korean in order to further test claims of cross-linguistic extensibility. Notably, SNACS has yet to be tested on a highly agglutinative language like Korean, whose adpositions (*josa*, §2.1) are bound morphemes suffixed on nominals, rather than independent lexical items like in English and Chinese. Korean adpositions are also peculiar in that some participate in case marking; we annotate postpositional nominative and accusative markers within the purview of SNACS.

Our contributions are three-fold: (1) we show that SNACS can be applied to Korean by adapting the SNACS hierarchy and guidelines to cover language-specific phenomena (§3); (2) we produce a broad-coverage corpus of Korean SNACS annotations and provide a corpus analysis of the Korean data (§4); and (3) we provide in-depth comparison between parallel Korean and English SNACS for purposes of our own study and for basis of comparison for future application of SNACS (§5).[1] Our work represents the first application of SNACS to an agglutinative language where adpositions are bound morphemes. Additionally, it represents a first Korean supersense corpus that was specifically produced for Korean postpositions.

[1] Korean SNACS guidelines and corpora are available at https://github.com/jdch00/k-snacs.

Proceedings of the 2nd International Workshop on Designing Meaning Representations, pages 53–66
Barcelona, Spain (Online), December 13, 2020

2 Background

2.1 Korean Postpositions

We focus on the well-researched category of *josa* (Sohn, 2001), as postpositions are known in Korean linguistics, as a target of our annotation. Characteristic of agglutinative languages, josa are bound morphemes that are suffixed on a nominal unit, though some pragmatically motivated postpositions may also attach to non-nominal units such as predicates and adverbs. As noted earlier, while many of them can be thought of as rough counterparts to English prepositions and hence act primarily as encoders of semantic relations, the functions carried out by josa run a broader gamut: some are case markers (nominative, accusative and genitive), while others supply pragmatic or contextual information.

In terms of syntactic distribution, josa have two noteworthy traits. First, some may be show up stacked, as exemplified in 1.[2] Such stacking is strictly governed by morphosyntatic rules. Another is that the case-marking josa are not mandatory: the nominative and accusative markers may be absent, leaving bare nominals in place (example 2a)—which is especially common in a spoken context where there is no ambiguity—or they may be superseded by another, pragmatically motivated, josa (2b).

(1) 나에게-만-이 아니라 우리 모두에게...
me-**DAT**-**FOC**-**NOM** not-but us all-DAT
"Not just to me but to all of us..."

(2) a. 빌이 점심(을) 먹었다
Bill-**NOM** lunch(-**ACC**) ate
"Bill ate lunch"

b. 빌은 점심(을) 먹었다
Bill-**TOP** lunch(-ACC) ate
"Bill ate lunch"

2.2 Related Work

Josa have received considerable theoretical attention within Korean linguistics. Much of the work has focused on investigating their syntactic function and patterns of grammaticalization, and enumerating prototypical semantics of specific groups of postpositions (e.g., Kang, 2012; Hwang, 2012; Sohn, 2001; Choi-Jonin, 2008; Rhee, 2004). Within semantics, postpositions have been investigated within semantic domains such as spatial configuration (e.g., Kang, 2012; Choi and Choi, 2018; Lee and Kabata, 2006). Little attention has been paid to establishing broader semantic categories of meaning that generalize over specific postposition types. While the morphosyntactic literature traditionally recognizes a dozen different grammatical categories (e.g., Nominative, Accusative, Dative, Genitive, Locative, Allative) for postpositions (e.g., Sohn, 2001), these josa categories, as this paper will show, are only partially adequate in the face of the full range of semantic behaviors we observe in the data. Semantically adequate and comprehensive annotation requires a richer palette of semantic labels that can apply broadly across the postpositional types, for which we turn to SNACS.

Computational approaches and resource creation projects have also attempted to classify Korean postpositions, with a focus on morphosyntax (in morphological tagging and syntactic parsing) (Choi and Palmer, 2011; Hong, 2009; Han, 2005). The Penn Korean Treebank (Han et al., 2001), for example, recognizes four part-of-speech (POS) categories (case, adverbial, conjunctive, and auxiliary) to cover all postpositional morphemes, and the 21st Century Sejong Project (Park and Tyers, 2019; Kim, 2006) retains a slightly larger inventory of nine POS tags, generally corresponding to the grammatical categories found in the traditional literature. More recently, the Korean Universal Dependency project guidelines do not directly address the individual postpositions since Korean postpositions are considered sub-lexical units. Instead, the POS category of `NOUN` is assigned to the full (noun + postposition) lexical unit (Oh et al., 2020; Chun et al., 2018).

The status of postpositions as functional categories also plays into the lack of specific attention in computational semantic resources. For example, while the labeling of Korean PropBank (Palmer et al., 2006) arguments is to some extent guided by the semantics of the postposition (e.g., a nominative marker might suggest ARG0 and a locative marker, ARGM-LOC), the labels are annotated at the lexical and phrasal level centering on nominal elements. In the case of Korean AMR (Choe et al., 2019), postpositions

[2]Here is a list of gloss abbreviations used in this paper: accusative (ACC), dative (DAT), focus (FOC), genitive (GEN), nominative (NOM), question (Q), speculation mood (SPEC) and information topic (TOP).

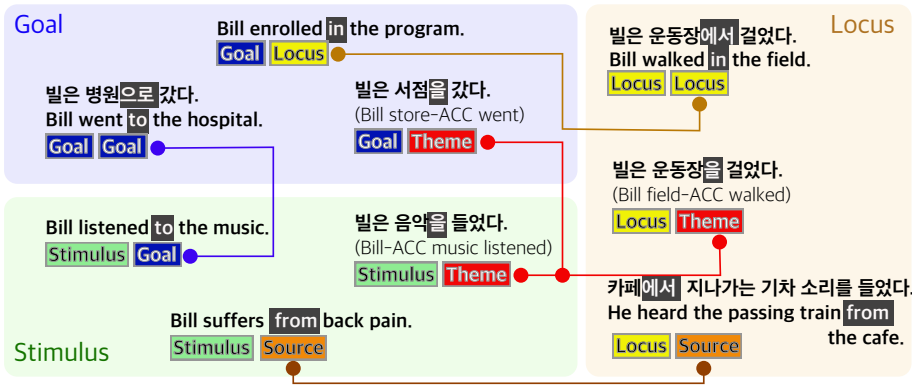

Figure 1: Conceptual diagram illustrating selected adposition usages and how they are related via construal within the SNACS framework. The light boxes cluster adpositional use by their scene roles. The lines indicate the connections various construals have via function labels—what the adposition lexically encodes. Some of these construals are found in both languages, others only in one or the other.

are omitted from the annotation entirely (as prepositions solely marking roles are omitted from English AMR; Banarescu et al., 2013).

SNACS is thus the most promising basis for a semantic framework with which to annotate Korean adpositions. The goal is to pave way for a full-scope, comprehensive treatment of the major semantic dimension of Korean postpositions. To the best of our knowledge, this work is the first to annotate a Korean corpus specifically targeting postpositions. Moreover, this work represents the first Korean application of a lexically-agnostic semantic analysis which cross-cuts adpositional types.

2.3 SNACS Framework

The Semantic Network of Adposition and Case Supersenses (SNACS; Schneider et al., 2018, 2020) is a framework for annotating adpositions with coarse-grained semantic classes called supersenses that broadly capture prepositional semantics without particular reference to any lexicon. The current version of the scheme defines 50 such supersenses for event participant roles (inspired by traditional thematic roles: AGENT, THEME, RECIPIENT, etc.), circumstantial roles (e.g. TIME, MANNER), and roles describing relationships between entities (e.g. POSSESSOR, WHOLE, QUANTITYVALUE). Annotating adposition uses in context serves to disambiguate them—e.g. "the wheel **of** the car" (WHOLE) versus "the destruction **of** the city" (THEME). Unlike dictionary senses (cf. Litkowski and Hargraves, 2005; Litkowski, 2014), the supersenses transcend lexical types in order to group together different adpositions with related meanings: thus WHOLE applies to both "the wheel **of** the car" and "the paint **on** the car". While the semantic criteria aim to be language-agnostic, the details of how to apply these labels to disambiguate adposition tokens in text—including specific criteria for which tokens to annotate, and how to deal with various language-specific constructions—need to be developed on a per-language basis. Extensive guidelines and multiple annotated corpora (web reviews; *The Little Prince*) are available for English. The Mandarin translation of *The Little Prince* has been fully annotated as well (Peng et al., 2020).

A distinctive aspect of SNACS is the so-called *construal analysis*, by which some tokens receive not one but two supersenses to reflect different facets of the usage: the **scene role** with respect to a larger situation (typically denoted by a predicate), and the **function** or primary lexical contribution of the adposition itself. These diverge in cases like "the paint **on** the car", where on the one hand the relationship between paint and car is one of part-whole; and on the other hand, the use of **on** frames it as a locational relationship. With the construal analysis this token would receive a scene role of WHOLE and function of LOCUS (WHOLE⇝LOCUS for short). By default, if the role and function are congruent, a single label is given.

3 Applying SNACS to Korean

We apply the SNACS supersenses to explicit mentions of Korean postpositions, and we find that the labels are generally applicable to Korean postpositional semantics. In this section, we discuss a few of the language-specific challenges we faced in applying SNACS to Korean.

Nominative and Accusative Cases One of the earliest challenges in adapting SNACS to Korean postposition semantics was deciding how to consistently label the nominative (NOM) and accusative (ACC) case markings. NOM and ACC case markings are postpositions that attach to the subject and object

of a sentence, respectively. Because these are identified in English via word order, there was no existing annotation scheme to follow in SNACS.

We adopt the view that the predicate's syntactic assignment of case marking such as NOM and ACC (as well as ergatives and absolutives) generally aligns with agentivity of the participant with respect to the verb given a transitive event (Grimm, 2011; Fillmore, 1968). Thus, for the function labels, we link the NOM and ACC labels to proto-agent and proto-patient roles (Dowty, 1991), respectively, in a transitive event: NOM (이/-i)[3] receives CAUSER or AGENT function label and ACC (을/-ul) receives the function of either THEME (i.e., general undergoer) or its subtype TOPIC (see example 3). In an intransitive event, where the NOM marks the patient argument, the function label of THEME is assigned (4).

For cases where the predicate assigns to NOM and ACC semantics that is different to that of their prototypical use, we represent the semantics assigned by the predicate as the scene role (e.g., ORIGINATOR of a communication event in 5 and LOCUS of an action in 6), while the function is the role associated most directly with the case marking (e.g. THEME for ACC). Our decisions are fully compatible with the treatment of English subjects and objects proposed in Shalev et al. (2019) (though available English SNACS corpora do not yet contain such annotations).

(3) 빌이/AGENT 사과를/THEME 먹었다
Bill-**NOM** apple-**ACC** ate
"Bill ate an apple"

(4) 해가/THEME 일찍 떴다
sun-**NOM** early rose
"the sun rose early"

(5) 빌이/ORIGINATOR⇝AGENT 대답했다
Bill-**NOM** answered
"Bill answered"

(6) 빌이/AGENT 공원을/LOCUS⇝THEME 걸었다
Bill-**NOM** park-**ACC** walked
"Bill took a walk in the park"

Contextual Postpositions. English SNACS has strictly focused on the annotation of semantic relations, excluding discourse connectives like "**according_to** him" or "**as_for** me" from annotation. We extend this treatment to two Korean discourse markers: vocative marker 야/-ya and politeness marker 요/-yo.

The pragmatic category in Korean, however, extends beyond these two markers. Korean also includes a category of frequently used pragmatic postpositions, whose role in a sentence is to evoke a particular set of contextual information regarding the entities to which they attach, thereby altering overall reading of the sentence. To address such usages, we introduce a new supertype CONTEXT as a fourth branch of SNACS hierarchy and add to it two new supersenses, TOPICAL and FOCUS.

We assign TOPICAL to postposition 은/-un that marks the information topic (TOP) in a sentence providing a contrast to a contextually available referent (7). FOCUS label is for postpositions that indicate the focus of a sentence (FOC), contributing information such as contrastiveness, likelihood, or value judgements (8 and 9).[4] There are a total of 10 identified postpositions that fall within the FOCUS category. In the example below, the three sentences have the same propositional value (i.e., "Bill did a good job"), but the postpositions situate the entity they mark within varying context.

(7) Bill은/TOPICAL 일을 잘 했다
Bill-**TOP** work-ACC well did
As for Bill, he did a good job.

(8) Bill만/FOCUS 일을 잘 했다
Bill-**only** work-ACC well did
Only Bill (and no one else) did a good job.

(9) Bill까지/FOCUS 일을 잘 했다
Bill-**even** work-ACC well did
Even Bill, the least likely candidate, did a good job.

Quotative Postpositions The Korean postposition inventory includes half a dozen markers that identify direct and indirect quotes in a sentence. The scene role that they play ranges from TOPIC (10), IDENTITY (11), to COMPARISONREF depending on the role assigned by the head verb. We propose a new supersense label QUOTE, a subcategory of THEME, to cover these at the function level.

(10) 오늘 도착한다고/TOPIC⇝QUOTE 했다
today arrive-**quote** say
"[They] said he'd arrive today."

(11) 지나를 천사라고/IDENT.⇝QUOTE 생각한다
Gina-ACC angel-**quote** thinks
"[They] thinks of Gina as an angel"

[3]NOM marker 가/-ka is an allomorphic variant of morpheme 이/-i, and ACC marker 를/-lul is an allomorph of the morpheme 을/-ul. A morpheme and its allomorphs are treated as a single postposition. Other morpheme-allophone pairs in this paper include: TOP marker 은/-un & 는/-nun, and goal/instrument marker 으로/-ulo & 로/-lo.

[4]We identify 7 sub-categories of FOCUS: contrast (translates roughly to *at the very least*), additive focus (*also*), exclusive focus (*only* in 8), negative polarity focus (*not even*), inclusive focus (*among others*), and two types of scalar focus (*merely*, *even* in 9). Since each postposition only maps to only one of these functions, we do not subdivide FOCUS according to use.

	Count		Count		Count
Documents (chapters)	27	Annotated P Targets	4166	Unique SNACS labels	42
Sentence	168	Unique Ps	39	Scene roles	41
Tokens	10939	Nominative & Accusative Ps	1676	Functions	32
Tokens w/explicit Ps	4020	Topical & Focus Ps	1319	Unique Construal pairs	108
		Construal Pairs: Scene = Function	3161	Scene = Function	31

Table 1: Statistics of the Korean *Little Prince* corpus.

The placement of the postposition under the THEME label was in recognition that these are participant arguments of verbs of communication (e.g., saying, telling) and cogitation (e.g., thinking, considering). What sets this usages apart from TOPIC is that by virtue of being marked by a QUOTE postposition, the sentence specifies that the information was heard or evidenced by the speaker/writer. Korean quotatives have also been widely studied as an evidential marker, which is not fully captured by its placement under the THEME. This is a topic of continued investigation.

Functionally Bleached Postposition 에/-ey Scholars have noted 에/-ey marks inanimate entities for spatial, temporal and goal type relations (Kang, 2012; Choi-Jonin, 2008; Rhee, 2004). The postposition, however, is a highly bleached one: its meaning is largely dependent on the sense assigned by the predicate. On its own, it simply serves to specify that the nominal is in a certain circumstantial relationship with the predicate. For this reason, we specify CIRCUMSTANCE at the function level, and let the scene role to disambiguate the relationship between predicate and the nominal[5].

(12) 깊숙한 곳에/LOCUS⇝CIRCUMSTANCE 보물을 감추고있는...
deep place-**ey** treasure hide
"a treasure hidden **in** a deep place"

(13) 나는 동이 틀 무렵에/TIME⇝CIRCUMSTANCE 우물을 발견했다
I-NOM sun-NOM rising cusp-**ey** well-ACC discovered
"I discovered a well **at** around the time of sun rise."

(14) 마음에/BENEFICIARY⇝CIRCUMSTANCE 좋은 말
heart-**ey** good words
"words that are good **for** the heart."

Postposition Stacking Because of the agglutinative nature of Korean grammar, postpositional markers can productively stack on top of each other as exemplified by (1). Postposition stacking is governed by grammatical rules and exhibits varying levels of grammaticalization (Schütze, 2001; Sohn, 2001). For SNACS, we consider only six stacked postpositions as a single unit: 에게서/-eykeyse, 한테서/-hantheyse, 에게로/-eykeylo, 에다(가)/-eyta(ka), 에서부터/-eyseputhe, and 으로부터/-uloputhe. These six have acquired noncompositional meanings. Otherwise, the stacked postpositions (as in example 1) are considered compositional and annotated as separate targets.

4 The Korean *Little Prince* Corpus

4.1 Data & Annotation

We annotate the Korean translation of Antoine de Saint-Exupéry's novella *The Little Prince* (어린 왕자),[6] which is available in various languages and has previously received attention from AMR annotation (Banarescu et al., 2013) and SNACS efforts for English and Chinese (Peng et al., 2020; Schneider et al., 2018). This corpus consists of 27 chapters with 10,939 tokens (table 1).

[5]Unlike other macrolabels like PARTICIPANT, CONFIGURATION, and CONTEXT (see figure 3 that are not directly used for annotation, CIRCUMSTANCE is used directly to annotate adpositions that contextualize a background setting or occasions for an event (e.g. "We drink eggnog **for** Christmas." Christmas is not so much *why* one might drink eggnog, rather the cirucumstance in which one might drink eggnog). By analogy, we are claiming here that 에/-ey sets the scene for the relationship between the verb and the marked nominal, and the nominal further specifies what that relationship is.

[6]The translation we use can be found at `http://cezz.com/blog/category/15`. This particular translation has been made freely available online by various sites for over a decade now. Unfortunately, the translator is unknown.

	All Postpositional Types			Only NOM, ACC, TOP, FOC			Excluding NOM, ACC, TOP, FOC		
	# Target	Scene	Function	# Target	Scene	Function	# Target	Scene	Function
Phase 1	1149	80.4%	87.2%	807	87.2%	90.5%	342	64.3%	79.5%
Phase 2	2465	83.8%	91.6%	1788	89.6%	93.4%	677	68.4%	87.0%
Overall	3614	82.7%	90.2%	2595	88.9%	92.5%	1019	67.0%	84.5%

Table 2: Inter-annotator agreement on the annotation of The Little Prince chapters 3,5-27.

Identifying postposition targets. We obtain automatic tokenization and morphological analysis via the KOMA tagger (Lee and Rim, 2009), which uses the morphological tagset from the Sejong Treebank (Hong, 2009). Postpositions are not treated as separate tokens. For each token the tagger analyzes internal morphological structure (e.g., the word token 영국의 is analyzed as 영국/NNP+의/JKG "England+GEN"). Target postpositions are identified by a subword morphological tag starting with J. In the case of stacking, a nominal can have more than one such postposition (e.g., 너/NP + 에게/JKB + 는/JX), in which case they are considered separate annotation targets.

Guidelines. In order to establish language-specific guidelines for Korean discussed in §3, we first selected the first three largest chapters in within the first chapters of the novella (1, 2, 4). The three chapters were double annotated by two linguists—an expert in Korean linguistics and a native Korean speaker, using the original SNACS guidelines established English. Standards for Korean SNACS (§3) were reached via analysis of disagreements during weekly discussions.

Annotation. Once the general guidelines were established, the remaining chapters (3, 5–27) were annotated by two linguists—the native Korean speaker who codeveloped the guidelines and a newly trained native speaker. The annotation was divided into two phases. In the first phase, the two annotators met on a weekly basis to discuss disagreements and tricky annotation cases. In the second phase, the annotators were given further independence to annotate without weekly discussions, and disagreements were tackled in two sessions: once at half-way point and once again at the end. All instances were double annotated this way and gold labels were adjudicated by consensus. Guidelines were also updated based on issues raised during the discussions.

4.2 Interannotator Agreement

Table 2 shows inter-annotator agreement rates for Chapters 3 and 5–27. On average, we observe rates of 82.7% on the scene role and 90.2% on the function label (first set of columns in table 2). Digging a little deeper, we uncover that the annotation of NOM, ACC, TOP and FOC markers (second set of columns in table 2; these vastly outnumber the other types) is much easier than the rest of the postpositional types (third set of columns in table 2). We expect that higher agreement is due the fact that there are only a limited number of supersenses available for these four postpositional types, especially for TOP, which categorically maps to the supersense TOPICAL. This also tells us that pragmatic uses, i.e., FOCUS, are clearly distinguishable for the native speakers.

Overall, agreement on function is higher than the scene agreement. This is expected as the assignment of scene depends more on the context of the postposition when compared with function, rather than the internal semantics of the postposition. It is also worth noting that despite the higher annotation targets and decreased discussion sessions, the agreements are higher in Phase 2, which suggests increased familiarity with the guidelines improves agreement.

4.3 Corpus Analysis

The Korean *Little Prince* corpus contains 4,020 tokens with one or more postpositions for a total of 4,166 postpositional targets. There are 42 unique SNACS labels and 108 unique construals represented in the data. Table 1 shows the full statistics of the annotated corpus.

NOM, ACC, FOC, TOP Postpositions. NOM, ACC, FOC, and TOP type postpositions account for over 70% of all annotation targets (table 1). The information topic postposition (은/-eun) is unambiguously

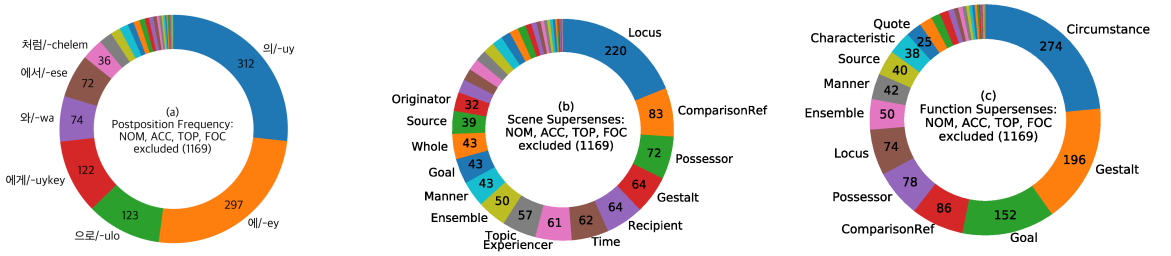

Figure 2: Distribution of postpositions and supersenses with nominative, accusative, information topic and focus postpositions are excluded. Distribution of the Korean postposition types in the corpus is in (a). Distributions of postposition supersenses at the scene role and function levels are shown in (b) and (c), respectively.

TOPICAL for both the scene role and the function, and all focus usages of postpositions (e.g., 도/-to, 만/-man) are labeled FOCUS. The most common supersenses for the NOM postposition (이/-i) are THEME (48.6%), AGENT (22.6%) and ORIGINATOR (19.5%) for the scene role, and THEME (52.0%) and AGENT (44.1%) for function. The majority of ACC supersenses are, unsurprisingly, THEME or its subtype TOPIC at both the scene role (83.1%) and function (82.2%) levels. We leave these four postpositional types out of the discussion for the remainder of §4.3.

Supersenses. Figures 2b and 2c shows the supersense label distributions for scene role and function. Out of the 42 unique SNACS labels, 15 labels are subtypes of CIRCUMSTANCE, 13 are CONFIGURATION subtypes, 11 are from PARTICIPANT, and 2 are from the new top-level category CONTEXT.

Eight supersenses in our inventory never appear in the annotated corpus. Four—INTERVAL, ORGROLE, SPECIES and STUFF—could be expressed postpositionally, though this never happened in the corpus. PURPOSE, APPROXIMATOR and COST are expressed through other grammatical categories such as verbal endings and nominal affixation,[7] which we do not currently annotate. The narrow set of usages listed in the English guidelines as TEMPORAL (as opposed to TIME, FREQUENCY, etc.) do not appear to correspond to any Korean postpositional usages, either.

Postposition Types. The frequency of Korean postpositions in the corpus is shown in figure 2a. What is interesting is that, in line with Croft (2000) and Dryer (1997), general linguistic categories assigned to Korean adpositions only partially describe their actual use. For example, 의/-uy, the most frequent postposition, is considered a genitive case marker in grammar texts (e.g., 빌의 집 "Bill**'s** house"), but it also exhibits a CHARACTERISTIC function that does not align with the genitive use like in 바둑판 무늬**의** 옷 "checkered-patterned-**uy** clothing" and a QUANTITYVALUE use exemplified by 세개**의** 화산 "three-**uy** volcanoes." In fact, it retains 18 distinct scene roles, which outnumbers the 16 unique scene roles for the bleached postposition 에/-ey (see §3), which we expected to be highly polysemous.[8]

The postpostion 으로/-ulo is another such case. It is primarily thought to mark DIRECTION (동쪽**으로** 가다 "go **towards** east"), INSTRUMENT (망치**로** 치다 "hit **with** a hammer"), and IDENTITY (바보**로** 여기다 "consider **as** dumb") (Kang, 2012; Choi-Jonin, 2008; Rhee, 2004). However, the data points to a wider variety of functions including MANNER (큰 소리**로** 떠들다 "make ruckus **in** a loud manner"), GOAL (학교**로** 가다 "go **to** school"), and MEANS (울음**으로** 깨웠다 "woke [me] up **by** (the means of) a loud cry"). In fact, 으로/-ulo is highly polysemous at 15 unique scene and 12 unique function labels.

5 Korean vs. English: An Inter-Annotation Discussion

We chose to annotate *The Little Prince* as it has been widely translated, and already annotated (partially) for SNACS in English (Schneider et al., 2018). In this section we compare chapters 1–7 of Korean and English *Little Prince* annotations.

[7]For example, the verbal ending 려고/-lyeko marks purpose in a sentential complement like 내가 먹으**려고** 샀다 "I bought it *for* eating", and the nominal suffix (distinct category from postpositions) 쯤/-ccum marks approximate time in 1시**쯤** "*about* 1 o'clock". These are currently under investigation for potential future addition to SNACS annotation.

[8]By way of comparison, Blodgett and Schneider (2018) applied SNACS to English possessives in online reviews. They report 15 supersenses as being attested for possessive pronouns/**'s**.

(a)

Korean targets: 1144 (447)	Uniq Ps: 29 (27)	Uniq scene: 36 (32)	Uniq functions: 27 (24)	Uniq construals: 75 (60)
English targets: 591	Uniq Ps: 60	Uniq scene: 45	Uniq functions: 39	Uniq construals: 97

(b)

Scene Roles				Functions				Construals			
KO		EN		KO		EN		KO		EN	
val	%	val	%	val	%	val	%	val	%	val	%
Theme	24.4	**Locus**	8.3	Theme	25.2	**Gestalt**	13.2	Topical↝Topical	24.2	**Topic↝Topic**	7.7
Topical	24.2	**Topic**	8.3	Topical	24.2	**Goal**	9.0	Theme↝Theme	24.1	**Locus↝Locus****	6.5
Focus	9.4	**CompRef**	5.4	Focus	9.4	Locus	9.0	Focus↝Focus	9.4	**Recipient↝Goal**	4.7
Locus	4.9	**Time**	5.1	Circums.	6.7	**Topic**	8.5	Agent↝Agent	4.4	Time↝Time	4.7
Topic	4.5	Recipient	4.9	Agent	5.9	**Possessor**	6.4	Stimulus↝Topic	3.1	Possessor↝Possessor	4.3
Agent	4.4	Manner	4.7	**Topic**	5.7	Source	6.3	**Locus↝Circums.****	2.8	Gestalt↝Gestalt	4.2
Stimulus	3.2	Whole	4.7	**Gestalt**	5.4	Time	5.1	**Topic↝Topic**	2.5	**CompRef↝CompRef**	4.2
CompRef	2.3	Gestalt	4.6	**Goal**	2.9	**CompRef**	4.9	**CompRef↝CompRef**	2.3	Source↝Source	4.0
Orig.	2.2	Possessor	4.4	**CompRef**	2.3	Identity	4.9	**Manner↝Manner**	1.6	Whole↝Gestalt	3.0
Time	2.0	Source	4.1	**Possessor**	2.2	Direction	3.6	**Recipient↝Goal**	1.6	**Manner↝Manner**	2.8

Table 3: (a) Distribution of adposition targets, supersenses and construal. For Korean numbers in parentheses specify counts when the NOM, ACC, and TOP postpositions are excluded. (b) A comparison between top 10 most frequent Korean and English scene roles, function labels and construals as found in the first 7 chapters of *The Little Prince*. The commonalities are marked in **bold**, Korean-only labels are highlighted in light gray, labels whose counts are highly influenced by NOM and ACC markers are in dark gray, and construals that are approximate cognates are marked with **.

5.1 Adposition, Supersense and Construal Distributions

Table 3 provides statistics and shows a side-by-side comparison between the top 10 most frequent scene roles, function labels, and construals in Korean and in English.

More Tokens, Fewer Types. We observe that given the translation of the same text, Korean postpositional targets outnumber English prepositions by nearly 2:1 in token count but are dwarfed in unique postposition types, about two-thirds that of English. This is mainly due to NOM, ACC and TOP, which attach to either subjects or objects of a verb as AGENT, THEME or TOPICAL. These account for over 60% (697 of 1144) of all targets, and they do not currently have annotation counterparts in the English annotation. Although the English SNACS project began applying supersenses to subjects and objects (Shalev et al., 2019), *The Little Prince* chapters do not yet reflect this update. The high token counts may also be related to the postpositional expression of FOCUS, which is currently unique to Korean.

Consequently, AGENT, THEME, and TOPICAL and FOCUS are the most frequent in the data. We observe two more scene labels used frequently by the NOM and ACC markers: ORIGINATOR (the source of communication; e.g., "*he* {said | replied}") and STIMULUS in the construal STIMULUS↝TOPIC (the object perception; e.g., "he saw *a picture of a boa*"). Interestingly, EXPERIENCER↝AGENTs, the experiencing counterparts of STIMULUS↝TOPICs (i.e., "*he* saw a picture of a boa") are much further down the frequency list, in large part due to the fact that Korean allows for subjects to be dropped if recoverable from context.

While attested adposition type counts are significantly lower than in English, Korean postpositions are more *polysemous* both in terms of scene roles and functions. In fact, over half of all postpositions are associated with two or more supersenses, while in English only about a third of the targets are associated with multiple supersenses.

Comparable Scene Roles and Diverging Functions & Construals. Overall, the two languages share the most frequent supersenses for scene roles. Beyond the supersenses shown to be in common in table 3, all of the top 10 English scene roles can be located among the list of top 15 Korean scene roles.

Label correspondences are lower at the function level, where 7 out of 10 most frequent English functions number among the top 15 Korean function supersenses. At first glance, it is admittedly odd that postpositions whose function is to deal with basic meanings like LOCUS, TIME, or BENEFICIARY should not number among the most frequent in Korean. However, this gap is explained by 에/-ey, a postposition associated with one function label, CIRCUMSTANCE (ranking 4th for function in Table 3; see §3 for

examples). With a single function label, it mediates 13 distinct scene roles in the first 7 chapters and 19 distinct scene roles in the whole of *Little Prince*.[9]

This agrees with what we saw earlier: Korean seemingly economizes on postposition types to express a variety of semantics. While at at the scene level Korean and English cover similar ground, English has a more diverse array of adposition choice. In order to gain a better understanding of the linguistic differences, we turn to an apples-to-apples comparison by aligning Korean and English annotations side-by-side to explore just how each language handles the same overall content.

5.2 Adposition Alignment Study

We pick the two chapters with the most English preposition targets, 2 and 7, to manually align the Korean and English annotations. Among 226 English and 410 Korean adpositions in these chapters, 81 were aligned based on the following criteria: firstly, the nominal to which Korean postposition is attached must refer to the same mention as the object of the English preposition, and secondly, the head [10] of the marked Korean nominal must refer to the same mention as the head of the English preposition. When possible, we align all markers including NOM and ACC markers (15), English possessives that internalize the object (i.e, my = of me) (16), and semantically approximate references like in 17 where "image" and "suggestion" are considered the same concept.

(15) 나를 바라보았다
me-**ACC** stare.
"[He] stared at me."
ALIGNED TO EN: He stared **at** me.

(16) 나는 얼마나 놀라웠겠는가
I-**TOP** just-how surprised-SPEC-Q
"Can you suppose just how surprised I was?"
ALIGNED TO EN: Imagine **my** amazement.

(17) 길 잃은 아이의 모습이 아니었다
way lost-TOP child-**GEN** image-NOM wasn't
"It wasn't an image of a lost child"
ALIGNED TO EN: Nothing about him gave any suggestion **of** a lost child

Among aligned adpositions, the two languages agree on scene role 66.7% of time and the function label agreement is 38.3%. While it is expected that the scene role should be higher in agreement than the function label as scene roles are assigned by the predicate or the verb, the numbers seem surprisingly low, especially for the function label. The intuition is, since two corpora represent parallel stories, we would expect the two languages to agree more at very least for the scene role. And this is not limited idiomatic usages that do not align like "I jumped to my feet" vs. "벌떡 일어섰다" (suddenly stood up) or partially align like in example 1 in table 4.

The best case disagreement scenario would be that of example 2 in table 4, where the same situation, thus same scene label, is mediated by differing postpositions according to the language-specific expectations (different functions); e.g., the STIMULUS of the staring event is realized with different adpositions in the two languages. These types of disagreement would account for low function label agreement, but this does not explain the scene roles.

What seems to be going on is that while Korean and English pairs do express parallel meaning via adpositions, the subject matter is handled in ways that the supersenses can't generalize. In some cases, the same semantics is conveyed from a different angle or point of view. For example, the third translation pair in table 4 is conceptually equal. But while English chooses to describe how long a flower has been producing thorns via DURATION semantics, Korean relays the same information by specifying STARTTIME of the event (which "The flowers have been making thorns **since** a million years ago"). We could possibly allow for SNACS to generalize the two sentences by stepping up the hierarchy to the TEMPORAL node. But there are other instances where generalization via hierarchy does not work so well.

Take for instance example 4, where English chooses MANNER, which is a part of CIRCUMSTANCE hierarchy and Korean chooses THEME from the PARTICIPANT tree. This difference is a direct result from

[9] Top 5 most frequent scenes for 에/-ey include LOCUS, TIME, GOAL, TOPIC and EXPLANATION in the full corpus.

[10] By "head" we mean the phrase to which the prepositional phrase (in English) or the marked nominal (in Korean) attaches in a constituency representation. In English, a head is most often a verb or a noun, and in Korean, most often the head is a verb (noun heads are possible but limited).

	English	Korean
1)	(surprised to see) a light break **over**/PATH the face [of the prince]	얼굴이/THEME 환하게 밝아지[다] *face-**NOM** brightly brighten-up*
2)	He stared **at**/STIMULUS⇝DIRECTION me thunderstruck	그는 어리둥절해서 나를/STIMULUS⇝THEME 바라보았다 *he-NOM puzzled me-**ACC** stared*
3)	The flowers have been growing thorns **for**/DURATION millions of years	수백만 년 전부터/STARTTIME 꽃들은 가시를 만들고 있어 *millions years prior-**since** flowers-TOP thorns-ACC make be*
4)	swell up **with**/MANNER pride	교만으로/THEME⇝MANNER 가득 차 있다 *I-**TOP** just-how pride-**with** full filled be*
5)	I was upset **over**/STIMULUS⇝TOPIC that bolt	나는 볼트 때문에/EXPLAN.⇝CIRCUMS. 신경이 곤두[섰다] *I-TOP bolt reason-**ey** nerves tensed-up*

Table 4: Examples of most common cross-linguistic differences among aligned adpositions. The aligned adpositions are in **bold** and the heads are underlined.

verb choice ("swell up" vs. "be filled"), which alters the scene role (i.e., how is it swelled up? vs. what is it filled with?). In example 5, the semantics of the head predicate are parallel, but the difference comes from a collocational difference based on which adposition the verb prefers. In the example, English chooses to talk about the topic of the pilot's emotion STIMULUS⇝TOPIC versus Korean's choice of an EXPLANATION modifier to describe the reason behind the pilot's mood (caused by the emotion).

These divergences are natural variations based on linguistic choice and are complicated by the fact that both English and Korean texts are translations of the original French novella. Bridging such translation divergences (Dorr, 1994; Deng and Xue, 2017) would require a richer modeling of causality and representations that will allow for deeper inferences about the divergent categories (Vyas et al., 2018; Hershcovich et al., 2019; Nikolaev et al., 2020; Briakou and Carpuat, 2020). We do not have a ready proposal to offer for bridging such differences through the SNACS framework. However, investigating further into nuanced semantics like casuality or force dynamics (Croft, 2015, 2012) that would aid generalizations certainly remains a compelling area of future research.

6 Conclusion & Future Work

In this work, we have presented the first annotated corpus of Korean preposition supersenses and included a detailed comparison of a subset of the data with a parallel English corpus. We find that, overall, Korean and English adpositions cover similar semantic ground, making English SNACS adaptable to typologically distant language like Korean. Still, applying the scheme to Korean required us to establish new supersenses for pragmatic usages not found in English adpositions, and to develop policies for case marking, among other innovations.

A number of directions remain for future work. One is an inquiry into supersenses that do not appear or appear infrequently in the annotated corpora (§4.3). We believe there may be certain language-specific phenomena at play (e.g., multi-word postpositions) meriting further investigation. The English-Korean parallel study also indicates a further need for investigating nuanced semantics like causality and force dynamics within the SNACS framework. Finally, as we have noted earlier, the SNACS framework has been recently applied to Mandarin Chinese (Peng et al., 2020). Preliminary adaptation efforts are also underway for Hindi (Arora and Schneider, 2020) and German (Jakob Prange and Nathan Schneider, personal communication). All three initiatives target *The Little Prince*. These efforts thus herald an auspicious opportunity for cross-linguistic comparison of adposition and case systems.

Acknowledgments

We thank Vivek Srikumar and Austin Blodgett for helpful discussions as we were formulating the details of Korean SNACS guidelines. We would also like to thank the anonymous reviewers for their insightful comments. This research was supported in part by NSF award IIS-1812778 and grant 2016375 from the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel.

References

Aryaman Arora and Nathan Schneider. 2020. SNACS annotation of case markers and adpositions in Hindi. Presented at SIGTYP 2020: The Second Workshop on Computational Research in Linguistic Typology.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proc. of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria.

Austin Blodgett and Nathan Schneider. 2018. Semantic supersenses for English possessives. In *Proc. of LREC*, pages 1529–1534, Miyazaki, Japan.

Eleftheria Briakou and Marine Carpuat. 2020. Detecting fine-grained cross-lingual semantic divergences without supervision by learning to rank. *arXiv:2010.03662 [cs]*.

Hyonsu Choe, Jiyoon Han, Hyejin Park, and Hansaem Kim. 2019. Copula and case-stacking annotations for Korean AMR. In *Proc. of the First International Workshop on Designing Meaning Representations*, pages 128–135, Florence, Italy.

Hong-yeol Choi and Youn Choi. 2018. 조사 '에','에서'의 공간의미 연구 - 인지의미론적 접근을 통한 조사의 의미자질 설정 가능성 고찰. *Yongbong Journal of Humanities*, 53:253–275.

Jinho D. Choi and Martha Palmer. 2011. Statistical dependency parsing in Korean: from corpus generation to automatic parsing. In *Proc. of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, pages 1–11, Dublin, Ireland.

Injoo Choi-Jonin. 2008. Particles and postpositions in Korean. *Typological Studies in Language*, 74:133.

Jayeol Chun, Na-Rae Han, Jena D. Hwang, and Jinho D. Choi. 2018. Building Universal Dependency treebanks in Korean. In *Proc. of LREC*, pages 2194–2202, Miyazaki, Japan.

William Croft. 2000. Parts of speech as language universals and as language-particular categories. In Petra M. Vogel and Bernard Comrie, editors, *Approaches to the Typology of Word Classes*, number 23 in Empirical Approaches to Language Typology, pages 65–102. De Gruyter Mouton, Berlin.

William Croft. 2012. *Verbs: Aspect and Causal Structure*. Oxford University Press, Oxford, UK.

William Croft. 2015. Force dynamics and directed change in event lexicalization and argument realization. In Roberto G. de Almeida and Christina Manouilidou, editors, *Cognitive Science Perspectives on Verb Representation and Processing*, pages 103–129. Springer International Publishing, Cham, Switzerland.

Dun Deng and Nianwen Xue. 2017. Translation divergences in Chinese–English machine translation: an empirical investigation. *Computational Linguistics*, 43(3):521–565.

Bonnie J. Dorr. 1994. Machine translation divergences: a formal description and proposed solution. *Computational Linguistics*, 20(4).

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Matthew S. Dryer. 1997. Are grammatical relations universal? In Joan L. Bybee, John Haiman, and Sandra A. Thompson, editors, *Essays on Language Function and Language Type: Dedicated to T. Givón*, pages 115–143. John Benjamins, Amsterdam.

Charles J. Fillmore. 1968. The case for case. In Emmon Bach and Robert Thomas Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart, and Winston, New York.

Scott Grimm. 2011. Semantics of case. *Morphology*, 21(3-4):515–544.

Chung-hye Han, Na-Rae Han, Eon-Suk Ko, Martha Palmer, and Heejong Yi. 2001. Penn Korean Treebank: Development and evaluation. In *Proc. of the 16th Pacific Asia Conference on Language, Information and Computation*, pages 69–78.

Na-Rae Han. 2005. Klex: A finite-state transducer lexicon of Korean. In *International Workshop on Finite-State Methods and Natural Language Processing*, pages 67–77. Springer.

Daniel Hershcovich, Omri Abend, and Ari Rappoport. 2019. Content differences in syntactic and semantic representation. In *Proc. of NAACL-HLT*, pages 478–488, Minneapolis, Minnesota.

Yun-Pyo Hong. 2009. 21 세기 세종 계획 사업 성과 및 과제 [21st Century Sejong Project results and tasks]. *새국어생활 [New Korean Life]*, 19(1):0–0.

Hwa-Sang Hwang. 2012. *국어 조사 의 문법 [Grammar of Korean Postpositions]*. Knowledge and Culture [지식 과 교양].

Jena D. Hwang, Archna Bhatia, Na-Rae Han, Tim O'Gorman, Vivek Srikumar, and Nathan Schneider. 2017. Double trouble: the problem of construal in semantic annotation of adpositions. In *Proc. of *SEM*, pages 178–188, Vancouver, Canada.

Yunkyoung Kang. 2012. *Cognitive linguistics approach to semantics of spatial relations in Korean*. Ph.D. thesis, Georgetown University, Washington, DC.

Hansaem Kim. 2006. Korean National Corpus in the 21st Century Sejong Project. In *Proc. of the 13th NIJL International Symposium*, pages 49–54. National Institute for Japanese Language Tokyo.

Do-Gil Lee and Hae-Chang Rim. 2009. Probabilistic modeling of Korean morphology. *IEEE transactions on audio, speech, and language processing*, 17(5):945–955.

Jeong-Hwa Lee and Kaori Kabata. 2006. A comparative cognitive-semantic analysis of spatial postpositions in Korean and Japanese. *담화와인지 [Discourse and Cognition]*, 13(2):187–203.

Ken Litkowski. 2014. Pattern Dictionary of English Prepositions. In *Proc. of ACL*, pages 1274–1283, Baltimore, Maryland, USA.

Ken Litkowski and Orin Hargraves. 2005. The Preposition Project. In *Proc. of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 171–179, Colchester, Essex, UK.

Dmitry Nikolaev, Ofir Arviv, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Maria Saeboe, and Omri Abend. 2020. Fine-grained analysis of cross-linguistic syntactic divergences. In *Proc. of ACL*, pages 1159–1176, Online.

Tae Hwan Oh, Ji Yoon Han, Hyonsu Choe, Seokwon Park, Han He, Jinho D. Choi, Na-Rae Han, Jena D. Hwang, and Hansaem Kim. 2020. Analysis of the Penn Korean Universal Dependency Treebank (PKT-UD): Manual revision to build robust parsing model in Korean. In *Proc. of IWPT*, pages 122–131, Online.

Martha Palmer, Shijong Ryu, Jinyoung Choi, Sinwon Yoon, and Yeongmi Jeon. 2006. Korean Propbank. Technical Report LDC2006T03, Linguistic Data Consortium, Philadelphia, PA.

Jungyeul Park and Francis Tyers. 2019. A new annotation scheme for the Sejong part-of-speech tagged corpus. In *Proc. of the 13th Linguistic Annotation Workshop*, pages 195–202, Florence, Italy.

Siyao Peng, Yang Liu, Yilun Zhu, Austin Blodgett, Yushi Zhao, and Nathan Schneider. 2020. A corpus of adpositional supersenses for Mandarin Chinese. In *Proc. of LREC*, pages 5988–5996, Marseille, France.

Seongha Rhee. 2004. Grammaticalization of spatio-temporal postpositions in Korean. *The Journal of Linguistic Science*, 31:169–188.

Nathan Schneider, Jena D. Hwang, Archna Bhatia, Vivek Srikumar, Na-Rae Han, Tim O'Gorman, Sarah R. Moeller, Omri Abend, Adi Shalev, Austin Blodgett, and Jakob Prange. 2020. Adposition and Case Supersenses v2.5: Guidelines for English. *arXiv:1704.02134v6 [cs]*.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proc. of ACL*, pages 185–196, Melbourne, Australia.

Carson T. Schütze. 2001. On Korean case stacking: The varied functions of the particles *ka* and *lul*. *The Linguistic Review*, 18(3):193–232.

Adi Shalev, Jena D. Hwang, Nathan Schneider, Vivek Srikumar, Omri Abend, and Ari Rappoport. 2019. Preparing SNACS for subjects and objects. In *Proc. of the First International Workshop on Designing Meaning Representations*, pages 141–147, Florence, Italy.

Ho-Min Sohn. 2001. *The Korean Language*. Cambridge University Press.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. In *Proc. of NAACL-HLT*, pages 1503–1515, New Orleans, Louisiana.

Additional Details

SNACS hierarchy. Figure 3 shows the hierarchy of 54 supersenses used for Korean SNACS annotation of *The Little Prince*. The pragmatic CONTEXT tree and its supertypes are new, as is the QUOTE under the PARTICIPANT tree. The scene role and function counts for each label in the corpus are shown in gray.

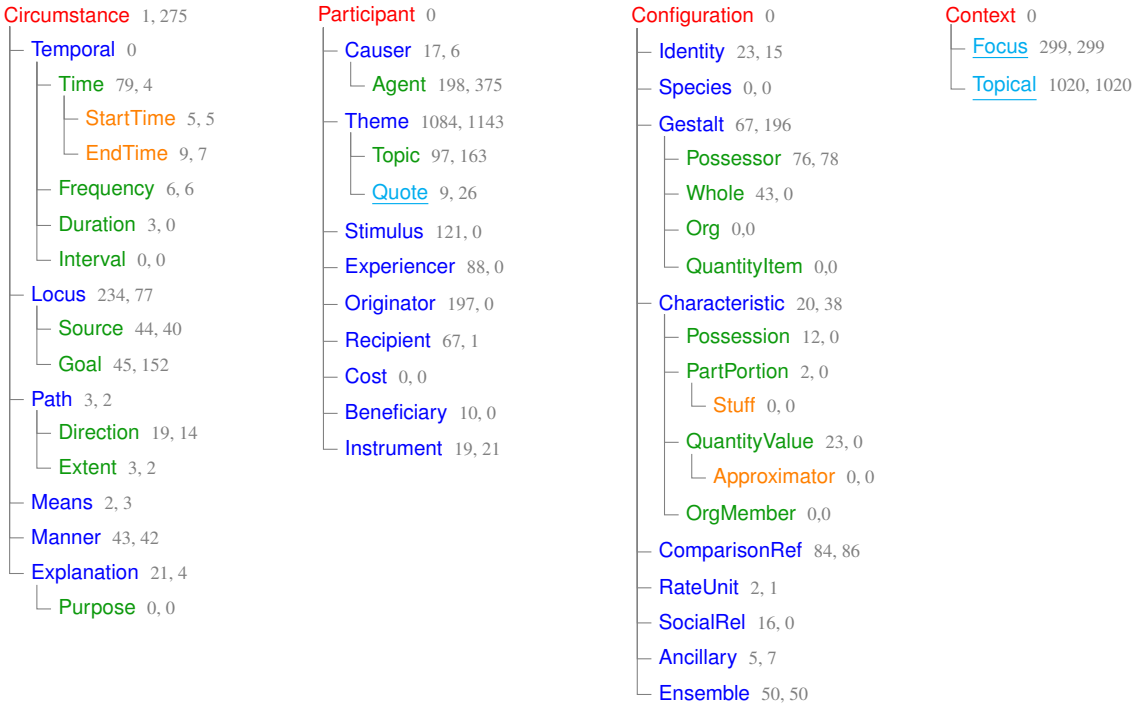

Figure 3: SNACS hierarchy of supersenses.