Abstract

We present an analysis of clausal nominalization developed in the context of the LinGO Grammar Matrix (Bender et al., 2002, 2010) to support the addition of subordinate clauses to the grammar customization framework. In particular, we examine the typological variation of nominalized clausal complements and nominalized clausal modifiers. To account for the range of variation in nominalized clauses across the world's languages and to support linguists in exploring alternative analyses, we propose a flexible library of analyses, allowing nominalization of the clause to occur at the V, VP or S level.

1 Introduction

Languages differ in the range of means they provide for expressing embedded propositions (propositions that serve as a dependent of some predicate). One prominent strategy in the world's languages is nominalization: a morphological or syntactic means of 'wrapping' a verbal constituent inside a nominal projection. This paper presents a cross-linguistic analysis of nominalized clauses in the context of a broader cross-linguistic grammar implementation project, namely the LinGO Grammar Matrix (Bender et al., 2002, 2010). The Grammar Matrix is a starter-kit for creating broad-coverage implemented precision grammars in HPSG (Pollard & Sag, 1994) which map between surface strings and Minimal Recursion Semantics (MRS; Copestake et al., 2005) representations. It includes a shared core grammar as well as a series of libraries extending that core with analyses for cross-linguistically variable phenomena. The analysis of nominalization presented here was developed in the context of our work on expressions of embedded propositions more generally, including as complements of verbs (Zamaraeva et al., to appear) and as modifiers of verbal projections (Howell & Zamaraeva, 2018). Typological surveys of these phenomena including Noonan 2007 show that clausal nominalization is a common strategy for embedded clauses, so we develop an analysis for nominalized clauses with these types of clausal subordination in mind.

As is typical for Grammar Matrix libraries, our analysis is intended to account for a broad range of typological possibilities as well as to give the user analytical freedom in modeling those possibilities. In particular, we allow for nominalization at different levels in the parse tree:

- Low: the nominal constituent is built out of a lexical verb (V)
- Mid: the nominal constituent is built out of a VP constituent comprising the verb and its complement
- High: the nominal constituent is built out of a full S: a verb plus all of its dependents

We also provide options on the semantic side, allowing high nominalization to be either strictly a syntactic phenomenon or one with semantic effects. A linguist

using the customization system can test alternative analyses in combination with analyses for other phenomena against text from their language to explore which best models the data.

We begin by describing in more detail the particular phenomena we are analyzing (§2) and briefly reviewing previous approaches (§3). We present our cross-linguistic analysis in §4, which includes the three levels of nominalization and two possible semantic representations. Finally we describe our implementation in the Grammar Matrix (§5) and how we evaluated the robustness of our analysis (§6). We conclude with a discussion of areas in which this work can be extended (§7).

2 Nominalized Subordinate Clauses

Nominalization is a common strategy for subordination in the world's languages (Noonan, 2007). To illustrate the difference between a verbal clause and a nominalized clause, consider the following data from Rukai, which contrasts the non-finite verb *amo-dhaace* 'leaving' in (1) with the nominalized *to'a-dhaac-ae* 'the reason for leaving' in (2).

(1) amo-dhaace =lrao
IRR-DYN.NFIN:leave =1SG.NOM
'I am leaving' [dru] (adapted from Zeitoun 2007)

(2) to'a-dhaac-ae =li
REAS.NMZ-DYN.NFIN:leave-REAS.NMZ =1SG.GEN
ma-lrakas-iae
STAT.FIN-dislike-1SG.OBL
'The reason why I'm leaving is because I dislike being here' [dru] (adapted from Zeitoun 2007)

In contrast with the non-nominalized form *amo-dhaace*, the nominalized verb *to'a-dhaac-ae* is marked with a nominalization circumfix which is specific to reason adverbial clauses. It also co-occurs with a genitive (rather than nominative) subject clitic. Nominalization morphemes and case frame change are common markers of nominalized clauses cross-linguistically, as shown in the following examples from Uzbek (3) and Irish (4). In fact, the Irish example demonstrates that case frame change for nominalized verbs is possible on on objects as well.[1]

(3) Xɔtin bu ɔdam-niŋ ǰoǰa-ni oğirla-š-i-ni istandi
woman this man-GEN chicken-OBJ steal-NMZ-3.SG-OBJ want.PST.3SG
'The woman wanted the man to steal the chicken.' [uzb] (adapted from Noonan 2007)

[1] In this example the subject is not overt in the nominalized clause. The genitive NP is the object.

(4) Is ionadh liom Seán a bhualadh Thomáis
COP surprise with.me John COMP hit.NMZ Thomas.GEN
'I'm surprised that John hit Thomas.' [gle] (adapted from Noonan 2007)

The characteristics of nominalized clauses in examples (2)–(3) may reflect the level at which nominalization occurred. The following examples of English gerunds (adapted from Malouf 2000) suggest a hierarchy of nominalization types for increasingly nominal properties of the phrase's internal distribution.

(5) a. The DA was shocked that Pat illegally destroyed the evidence.
b. The DA was shocked that she illegally destroyed the evidence.

(6) a. The DA was shocked by Pat having illegally destroyed the evidence.
b. The DA was shocked by her having illegally destroyed the evidence.

(7) a. The DA was shocked by Pat's having illegally destroyed the evidence.
b. The DA was shocked by her having illegally destroyed the evidence.

(8) a. The DA was shocked by Pat's illegal destroying of the evidence.
b. The DA was shocked by her illegal destroying of the evidence.

(9) a. The DA was shocked by Pat's illegal destruction of the evidence
b. The DA was shocked by her illegal destruction of the evidence
(adapted from Malouf 1998)

Malouf (1998) notes that (5) has no internal properties of an NP and is a fully verbal phrase: the *destroyed* is modified by an adverb and its subject's and object's case markings are consistent with those of English verbs. On the other hand, (9) has all of the properties of an NP and is a deverbal noun: *destruction* is modified with an adjective and its subject and object are both marked with different cases than those of the verb in (5).[2] The remaining examples illustrate the range between fully verbal and fully nominal expressions.

We take this variation in verbal and nominal properties as an indication of the level at which the verbal projection took on the properties of nominal projections, or put another way, at what level the clause was nominalized. In §4, we propose an analysis based on this observation, such that high nominalization (at S) allows adverbial modifiers and does not allow case change on subjects or objects; mid nominalization (at VP) allows adverbial modifiers and only allows case change on subjects; and low nominalization (at V) allows adjectival modifiers and case change on both subjects and objects.

3 Previous Approaches

Malouf (1998) provides a thorough review of previous approaches to clauses with both nominal and verbal characteristics. Here, we summarize his review as well as his own approach in order to situate our analysis within this body of work.

[2]Here we take *of* to be a kind of case-marking preposition.

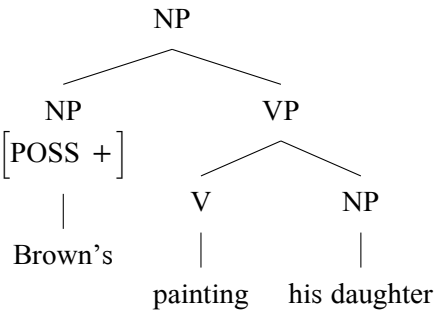


Figure 1: Pullum's approach (Malouf, 1998)

Pullum (1991) presents an analysis for English gerunds in which the VP headed by the verbal gerund combines with a possessive NP to form a larger NP constituent, the nominalized clause, as illustrated in figure 1. Lapointe (1993), on the other hand, takes a different approach, proposing a dual lexical category $\langle X|Y \rangle$ such that X determines the external distribution and Y the internal structure. Thus in the case of gerunds or nominalized clauses, the underlying lexical type would be $\langle N|V \rangle$. Malouf notes that neither approach accounts for gerunds with accusative subjects, e.g. *her having illegally destroyed the evidence* in (6) or adjective modification, as in *my wicked leaving my father's house*, as seen in old English. Furthermore, while Pullum's approach violates the principle of endocentricity by positing a head daughter which does not have the same distribution as the phrase, Lapointe's approach could generalize to other mixed categories that do not occur in the worlds languages.

Bresnan (1997) proposes a 'change-over' approach, wherein the verbal constituent changes to a nominal constituent, as illustrated in figure 2. In doing so, the gerund will have the properties of a verb until the change over occurs, and then will take on the properties of a noun. Malouf notes that in addition to violating the principle of endocentricity like Pullum's approach, this analysis also doesn't correctly account for adverb position. In particular, the gerund is the daughter of NP, so an adverb would attach after the gerund, not before. This incorrectly predicts *Pat's watching avidly movies* and incorrectly rules out *Pat's avidly watching movies*.

Finally, Malouf (1998) presents a mixed category analysis, positing a gerund head value, modeled with multiple inheritance, as shown in the hierarchy in figure 3. This allows gerunds to interact with phrase structure rules sometimes like verbs and sometimes like nouns. He pairs this with a lexical rule that derives the valence properties of the gerunds and shows how a similar approach can work for an variety of languages, including English, Arabic, Boumaa Fijian, and Dagaare.

Malouf argues against 'change-over' approaches (e.g. that of Bresnan 1997, inter alia), because they don't constrain what kinds of change overs are available.

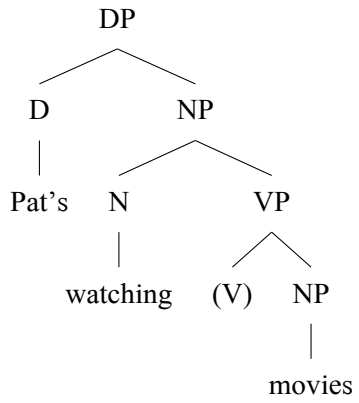


Figure 2: Bresnan's change-over approach (Malouf, 1998)

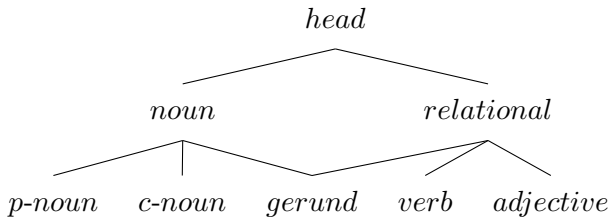


Figure 3: Malouf's multiple inheritance hierarchy (Malouf, 1998)

His mixed-category approach, combined with language-specific versions of the Head-Specifier and Head-Subject rules, elegantly accounts the mixed behavior of verbal gerunds. However, given the goal of grammar customization and the context of the Grammar Matrix code base, we take a change-over approach as it integrates more easily with the other libraries providing the phrase structure rules. On our analysis, the change over can happen at the S, VP or V levels. High nominalization (at S) allows adverbial modifiers and does not allow case change on subjects; mid nominalization (at VP, as in (7)) allows adverbial modifiers and only allows case change on subjects; and low nominalization (at V, as in (8)) allows adjectival modifiers and case change on both subjects and objects.[3]

In the next section, we present an analysis akin to that of Bresnan 1997, in that we take a change-over approach, using unary rules to transform verbal projections in to nominal projections. It differs in that it also includes lexical rules. Accordingly the change over of HEAD value need not correlate with the changes to constraints on arguments. This avoids some of the problems that Malouf (1998) finds with change-over approaches.[4] Acknowledging that this approach violates the principle of endocentricity, our goal in the Grammar Matrix is to facilitate modeling grammars, rather than narrowing the class of possible languages. Furthermore, we find that this change-over approach allows us to account for case-frame changes as well as adjective and adverb attachment effectively, in order to model the range of nominalization strategies discussed in the previous section.

4 A Cross-linguistic Analysis

In this section we present three distinct analyses for nominalization to account for the variation described in §2. We begin by introducing the NMZ feature in §4.1. This is followed by a description of three analyses for high (§4.2), mid (§4.3) and low (§4.4) nominalization, which we motivate using the data presented in §2.We discuss the additional work necessary to accommodate case frame changes in §4.5 and propose two possible semantic representations of nominalized clauses in §4.6.

4.1 The NMZ feature

Our analysis allows for the disassociation of the nominalization morphology from the actual change of the HEAD value from *verb* to *noun*. To facilitate this, we propose a Boolean HEAD feature NMZ, which we use to distinguish verbs inflected with a nominalization morpheme (but not yet nominalized) from other verbs. We also use this feature to differentiate between nominal constituents built from nominalized verbs and other (lexical) nouns. Nouns and verbs in the lexicon are constrained to be [NMZ $-$] and this constraint is changed to [NMZ $+$] only by

[3] Neither (5) nor (9) involve nominalization of the type we are concerned with; the former because the constituent is verbal at all levels, and the latter because the clause has no verbal properties.

[4] We leave to future work the project of ensuring that our analysis can account for all the data presented in Malouf 1998.

nominalization lexical rules. The low nominalization analysis changes the HEAD value from *verb* to *noun* in the lexical rule. However, the mid and high nominalization analyses employ a unary rule to change the HEAD value and that unary rule has [NMZ +] on both the daughter and mother. These processes are illustrated in detail in figures 4–6 below. Under our analysis, complementizers, subordinators and clausal verbs that require nominalized clausal complements constrain their complement to be both [NMZ +] to prevent selection of a lexical noun and [HEAD *noun*] to prevent selection of a verb that has gone through the lexical rule, but not the corresponding unary rule.

4.2 High Nominalization

Our first nominalization analysis involves nominalization at the S level, such that the constituent maintains verbal properties until all arguments are picked up (including the subject) and only then is the nominal constituent built. We have not found clear evidence that this option is attested in the world's languages: such evidence would involve a case language with nominalized clauses and no case change on the subject. Nevertheless, we provide this analysis as an option to linguists who may wish to test it against their data.

To accommodate clauses that remain verbal until all valence features are satisfied and then undergo nominalization, we posits two rules: a lexical rule that puts a morpho-syntactic marker on the verb and a unary phrase structure rule that builds a nominal constituent out of a verbal one. This is illustrated with the hypothetical example *Pat destroying the evidence*, where we pretend that Pat is a nominative subject (contrary to the facts of English), in figure 4.

The lexical rule is shared with the analysis for mid nominalization (§4.3), and accordingly is named *high-or-mid-nominalization-lex-rule*. This rule, defined in (10), adds [NMZ +] to the mother and identifies the INDEX of the daughter's subject with the INDEX of the mother's subject. We constrain only the subject's INDEX in order to accommodate case change under the mid-nominalization analysis. However, for high nominalization, a sub-type of this rule identifies the entire subject between the mother and daughter.[5]

(10)
$$\begin{bmatrix} \textit{high-or-mid-nominalization-lex-rule} & \\ \text{SYNSEM} \mid \text{LOCAL} & \begin{bmatrix} \text{CAT} \mid \text{HEAD} & \begin{bmatrix} \text{NMZ} & + \end{bmatrix} \\ \text{VAL} \mid \text{SUBJ} & \langle\, \text{INDEX}\, \boxed{0} \,\rangle \end{bmatrix} \\ \text{DTR} \mid \text{SYNSEM} \mid \text{LOCAL} & \begin{bmatrix} \text{CAT} \mid \text{VAL} \mid \text{SUBJ} & \langle\, \text{INDEX}\, \boxed{0} \,\rangle \end{bmatrix} \end{bmatrix}$$

Once the morpho-syntactic marker NMZ + has been added to the verb and its

[5]The AVMs shown in this paper are abbreviated in order to focus on features of interest. The lexical rules produced by the Grammar Matrix customization system also have many constraints that serve to copy information from daughter to mother. The reader can assume that all features are copied from daughter to mother unless otherwise specified. Grammars that exemplify these constraints can be checked out from revision 41825 here: svn://lemur.ling.washington.edu/shared/matrix/trunk/gmcs/regressiontests

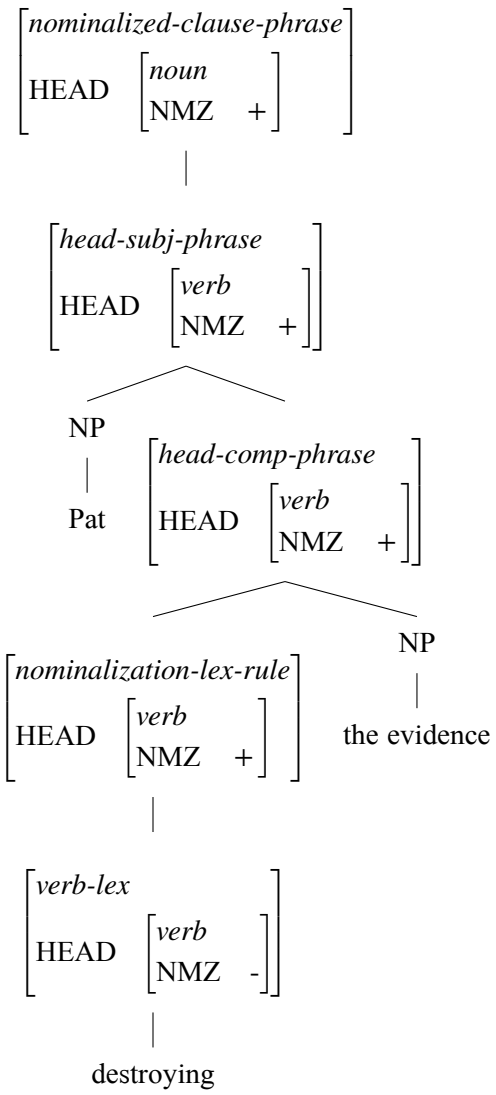
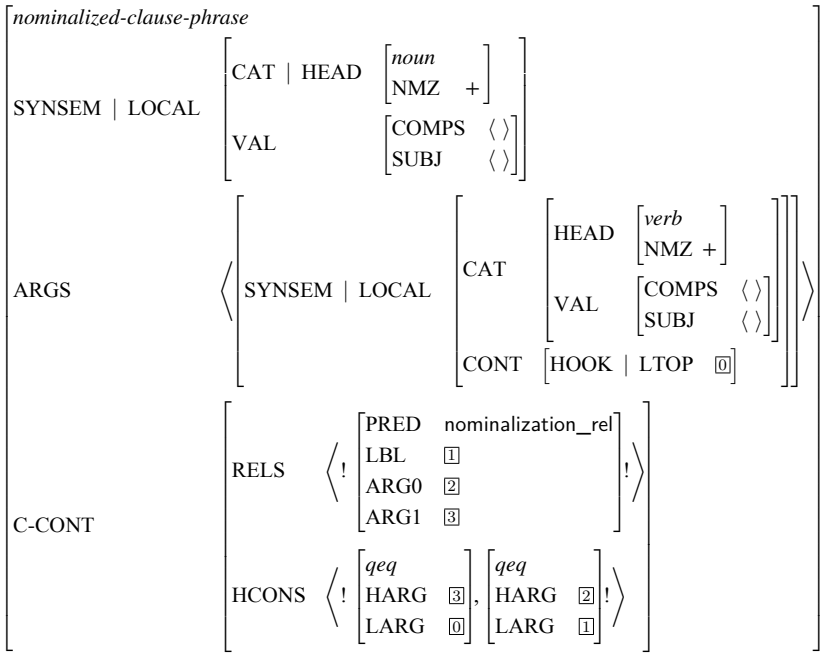


Figure 4: High nominalization

valence requirements have been satisfied, the clause can serve as the daughter of the *nominalized-clause-phrase* unary rule, defined in (11).

(11)


We constrain both the SUBJ and COMPS lists to be empty on the mother and daughter, so that this rule will only select clauses which are valence saturated. This rule effects the syntactic change from verbal to nominal projection, changing the HEAD type to *noun*. The unary rule also adds the necessary semantic constraints for the nominalized verb to be represented as a noun. This is accomplished by adding nominalization_rel to the C-CONT (constructional-content) list and linking the ARG1 of that predication to the daughter's LTOP.[6] This has the effect of 'wrapping' a nominal predication around the proposition built by the verb. The resulting MRS representation will be discussed in more detail in §4.6.

4.3 Mid Nominalization

Our next analysis involves the nominalization of verb phrases, i.e. verbal projections with empty COMPS lists. This analysis is motivated by examples such as (6) and (7), repeated here as (12) and (13).

(12) a. The DA was shocked by Pat having illegally destroyed the evidence.
b. The DA was shocked by her having illegally destroyed the evidence.

(13) a. The DA was shocked by Pat's having illegally destroyed the evidence.

[6]This connection is mediated by an 'equal modulo quantifiers' constraint (*qeq*) given in the value of HCONS. These constraints are part of the MRS analysis of quantifier scope ambiguity (Copestake et al., 2005) and introducing one here allows quantifiers in the nominalized clause to have the option of scoping below the embedding predicate, as desired.

b. The DA was shocked by her having illegally destroyed the evidence.

These examples exhibit hybrid properties: In (12) and (13) the verb is modified by an adverb and its complement bears its canonical case, i.e. within the VP constituent we see verbal properties. However, the subject appears with a non-canonical case, genitive or accusative.

Our mid nominalization analysis is very similar to the high nominalization analysis in that a morpho-syntactic marker is added by the *high-or-mid-nominalization-lex-rule* and the projection is changed from verbal to nominal by a unary rule higher in the tree, as illustrated in figure 5.

The lexical rule in (10) is also used for mid nominalization. As discussed in the previous section, this rule only identifies the INDEX of the subject, allowing the case value of the subject to be changed.[7] This process is described in more detail in §4.5. This analysis also uses a unary rule change the projection from verbal to nominal. The *mid-nominalized-clause-phrase* rule in (14) differs from the rule in (11) in only one way: instead of an empty subject list, the subject list of the daughter is constrained to be non-empty and identified with the the subject list of the mother.

(14)
$$\left[\begin{array}{ll}\textit{mid-nominalized-clause-phrase} & \\ \text{SYNSEM} \mid \text{LOCAL} & \left[\begin{array}{ll}\text{CAT} \mid \text{HEAD} & \left[\begin{array}{ll}\textit{noun} & \\ \text{NMZ} & + \end{array}\right] \\ \text{VAL} & \left[\begin{array}{ll}\text{COMPS} & \langle\,\rangle \\ \text{SUBJ} & \boxed{0} \end{array}\right] \end{array}\right] \\ \text{ARGS} & \left\langle \left[\begin{array}{ll}\text{SYNSEM} \mid \text{LOCAL} & \left[\begin{array}{ll}\text{CAT} & \left[\begin{array}{ll}\text{HEAD} & \left[\begin{array}{ll}\textit{verb} & \\ \text{NMZ} & + \end{array}\right] \\ \text{VAL} & \left[\begin{array}{ll}\text{COMPS} & \langle\,\rangle \\ \text{SUBJ} & \boxed{0} \end{array}\right] \end{array}\right] \\ \text{CONT} & \left[\text{HOOK} \mid \text{LTOP} \quad \boxed{1}\right] \end{array}\right] \end{array}\right] \right\rangle \\ \text{C-CONT} & \left[\begin{array}{ll}\text{RELS} & \left\langle ! \left[\begin{array}{ll}\text{PRED} & \text{nominalization\_rel} \\ \text{LBL} & \boxed{2} \\ \text{ARG0} & \boxed{3} \\ \text{ARG1} & \boxed{4} \end{array}\right] ! \right\rangle \\ \text{HCONS} & \left\langle ! \left[\begin{array}{ll}\textit{qeq} & \\ \text{HARG} & \boxed{4} \\ \text{LARG} & \boxed{1} \end{array}\right], \left[\begin{array}{ll}\textit{qeq} & \\ \text{HARG} & \boxed{3} \\ \text{LARG} & \boxed{2} \end{array}\right] ! \right\rangle \end{array}\right] \end{array}\right]$$

[7]Under our analysis mid nominalization without case change is allowed. While it is typologically unlikely that a language would have VP nominalization without case change on the subject (hypothetically exemplified by an adjective modifier above VP but below the subject), it is possible that a user developing a grammar for a language without a case system would want to avoid adding the additional case-change-related constraints to their grammar.

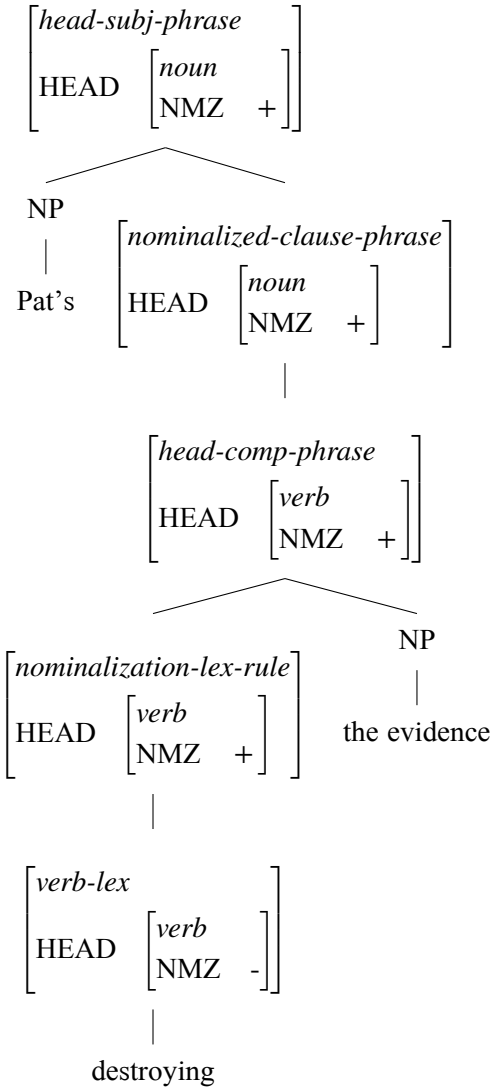


Figure 5: Mid nominalization

4.4 Low Nominalization

Our final analysis involves nominalization at the lexical level, before the underlying verb combines with any of its arguments. Although this analysis for nominalization occurs on the lexical level, we do not claim that it extends to all deverbal nouns. In particular, it is only appropriate for productive morphology which furthermore results in event nominalization (as opposed to e.g. agent nominalization). This analysis is appropriate for examples where the nominalized verb is modified by a low-attaching adjective and/or the case on the verb's complement differs from that found in its ordinary (non-nominalized) use. (8), repeated here as (15), falls into this category:

(15) a. The DA was shocked by Pat's illegal destroying of the evidence.
b. The DA was shocked by her illegal destroying of the evidence.

It may be that low nominalization is also motivated by changes to the CASE or HEAD value required of the complement. Under our analysis, these are actually always handled low (in the lexical rule), but linguists may prefer to analyze them as co-incident with the change of the HEAD and INDEX on the nominalized constituent itself.

Under our analysis of low nominalization, the lexical rule that provides the nominalization morpheme and the morpho-syntactic marker also directly changes the verb to a noun, as illustrated in figure 6. This rule, shown in (16), specifies [HEAD *noun*] and [NMZ +] on the mother. The lexical rule also adds the predication nominalization_rel to the MRS and links its first argument with the daughter (via a *qeq* constraint).

(16)
$$\left[\begin{array}{ll} \textit{low-nominalization-lex-rule} & \\ \text{SYNSEM} \mid \text{LOCAL} & \left[\text{CAT} \left[\begin{array}{l}\text{HEAD}\begin{bmatrix}\textit{noun}\\ \text{NMZ} \quad +\end{bmatrix}\\ \text{VAL} \mid \text{SUBJ} \langle\, \text{INDEX} \quad \boxed{0} \,\rangle\end{array}\right]\right] \\ \text{DTR} \mid \text{SYNSEM} \mid \text{LOCAL} & \left[\begin{array}{ll}\text{CAT} & \left[\text{VAL} \mid \text{SUBJ} \langle\, \text{INDEX} \quad \boxed{0} \,\rangle\right]\\ \text{CONT} & \left[\text{HOOK} \mid \text{LTOP} \quad \boxed{1}\right]\end{array}\right] \\ \text{C-CONT} & \left[\begin{array}{ll}\text{RELS} & \left\langle !\begin{bmatrix}\text{PRED} & \textsf{nominalization\_rel}\\ \text{LBL} & \boxed{2}\\ \text{ARG0} & \boxed{3}\\ \text{ARG1} & \boxed{4}\end{bmatrix}! \right\rangle \\ \text{HCONS} & \left\langle !\begin{bmatrix}\textit{qeq}\\ \text{HARG} \quad \boxed{4}\\ \text{LARG} \quad \boxed{1}\end{bmatrix}, \begin{bmatrix}\textit{qeq}\\ \text{HARG} \quad \boxed{3}\\ \text{LARG} \quad \boxed{2}\end{bmatrix}! \right\rangle\end{array}\right] \end{array}\right]$$

The lexical rule in (16) is a somewhat underspecified supertype that is further constrained depending on the specifications given by a user for a particular

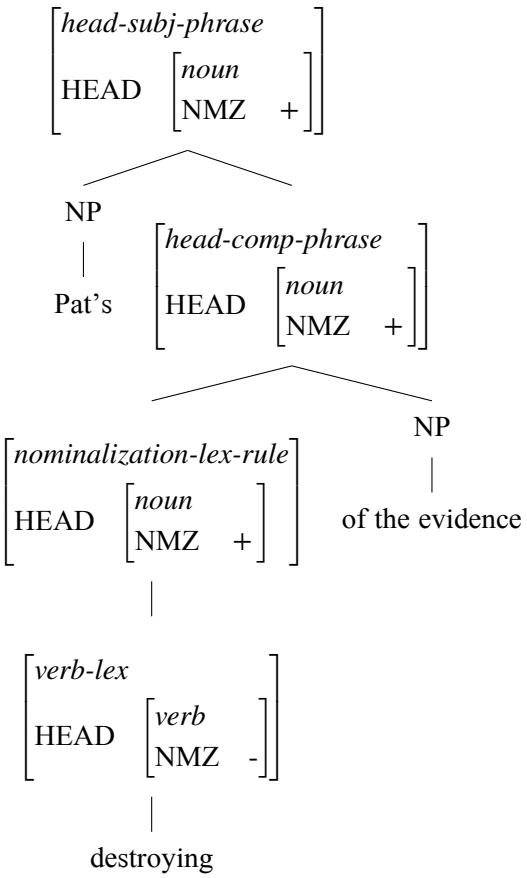


Figure 6: Low nominalization

language. It identifies the INDEX of the mother's subject with the INDEX of the daughter's subject. If the case on the subject changes upon nominalization, this constraint is sufficient (in combination with constraints on case discussed in §4.5 below). However, if case frame change does not occur, then we create a subtype of this rule that identifies the entire subject, rather than just the INDEX. Similarly, we add constraints to subtypes of this rule based on whether or not the object's case is changed. If the case on the object changes, a constraint to identify the complement's INDEX[8] between mother and daughter is added, whereas if the object's case does not change or the verb is intransitive, the entire complements list is identified between daughter and mother.[9]

4.5 Accommodating Case Frame Changes

In §4.3 and §4.4 we noted that the nominalization lexical rule supertypes only identify the indices of subjects and complements and that work remains to be done if the case frame of the nominalized verb differs from that of a non-nominalized verb. Subtypes of these rules are used to make changes to the HEAD features, including both the case and the associated head type.

In particular, when a user of the Grammar Matrix defines a morphological rule associated with nominalization, they may also indicate the case of the subject and/or object if they differ from the standard verbal case-frame. These case constraints are then added to the nominalization lexical rules by the Grammar Matrix customization system. Because certain cases may be associated with particular HEAD types in the language, the customization system has built-in functions for detecting the head types that are compatible with given case. We use these functions to identify the appropriate HEAD type and add that constraint to the lexical rule as well. Thus a hypothetical language in which nominalized verbs require genitive subjects and genitive case is marked by a preposition would have the following rule, inheriting from the *low-nominalization-lex-rule*.

(17)
$$\left[\begin{array}{l} \textit{low-intransitive-nominalization-lex-rule} \\ \text{SYNSEM} \mid \text{LOCAL} \mid \text{CAT} \mid \text{VAL} \left[\begin{array}{ll} \text{SUBJ} & \left\langle \left[\begin{array}{l} \text{INDEX } \boxed{0} \\ \text{HEAD} \left[\begin{array}{l} \textit{prep} \\ \text{CASE gen} \end{array}\right] \end{array}\right] \right\rangle \\ \text{COMPS} & \langle\,\rangle \end{array}\right] \\ \text{DTR} \mid \text{SYNSEM} \mid \text{LOCAL} \mid \text{CAT} \mid \text{VAL} \left[\begin{array}{ll} \text{SUBJ} & \langle\, \text{INDEX } \boxed{0} \,\rangle \\ \text{COMPS} & \langle\,\rangle \end{array}\right] \end{array}\right]$$

[8] Currently ditransitive verbs are not supported by the Grammar Matrix, so our analysis only accounts for one complement.

[9] For languages with case change on the object, we use two separate rules, one for transitive verbs which identifies the object's INDEX and one for intransitive verbs that identifies the entire COMPS list.

4.6 Semantic Representations

We provide two possible representations for nominalized clauses, using Minimal Recursion Semantics (MRS; Copestake et al., 2005). On the one hand, in many languages it can be argued that a nominalized subordinate clause has a different meaning than a fully verbal subordinate clause. At the very least, there must be a nominal predication in the semantic representation to which adjectives, like that in (8), repeated here as (18), and quantifiers can attach.

(18) a. The DA was shocked by Pat's illegal destroying of the evidence.
b. The DA was shocked by her illegal destroying of the evidence.

On the other hand, a linguist modeling a language in which nominalization is the only strategy for subordination might argue that there is no difference in meaning between nominalized subordinate clauses in that language and subordinate clauses in other languages. Therefore, we provide both options for our high nominalization analysis: one with a nominalization_rel and one without. At this time we do not allow a representation without a nominalization_rel for low and mid nominalization as this would prevent adjective modification of those clauses. This option may be appropriate to add, but only in languages which never allow adjectival modification of the low or mid nominalized structures.

For an example like the Turkish sentence in (19) with a nominalized clausal complement, the analyses described earlier in this section result in the MRS semantic representation in (20).[10]

(19) senin sinema-ya gel-me-n-i isti-yor-um
2SG.GEN cinema-DAT come-NMZ-2SG-ACC want-PROG-1SG
"I want you to come to the movies." [tur] adapted from Kornfilt (1997, p. 48)

(20)
$$
\begin{array}{l}
\langle\, h_1, e_2, \\
\quad \left| \begin{array}{l}
h_3{:}\mathsf{pron\_rel}(\textsc{arg0}\ x_4), \\
h_5{:}\mathsf{exist\_q\_rel}(\textsc{arg0}\ x_4,\ \textsc{rstr}\ h_6,\ \textsc{body}\ h_7), \\
h_8{:}\mathsf{\_cinema\_n\_rel}(\textsc{arg0}\ x_9), \\
h_{10}{:}\mathsf{exist\_q\_rel}(\textsc{arg0}\ x_9,\ \textsc{rstr}\ h_{11},\ \textsc{body}\ h_{12}), \\
h_{13}{:}\mathsf{come\_v\_rel}(\textsc{arg0}\ e_1,\ \textsc{arg1}\ x_4,\ \textsc{arg2}\ x_9), \\
h_{15}{:}\mathsf{nominalization\_rel}(\textsc{arg0}\ x_{17},\ \textsc{arg1}\ h_{16}), \\
h_{18}{:}\mathsf{exist\_q\_rel}(\textsc{arg0}\ x_{17},\ \textsc{rstr}\ h_{19},\ \textsc{body}\ h_{20}), \\
h_{23}{:}\mathsf{pron\_rel}(\textsc{arg0}\ x_{22}), \\
h_{24}{:}\mathsf{exist\_q\_rel}(\textsc{arg0}\ x_{22},\ \textsc{rstr}\ h_{25},\ \textsc{body}\ h_{26}), \\
h_{21}{:}\mathsf{want\_v\_rel}(\textsc{arg0}\ e_2,\ \textsc{arg1}\ x_{22},\ \textsc{arg2}\ x_{17})
\end{array} \right| \\
\quad \{\, h_6 =_q h_3,\ h_{11} =_q h_8,\ h_{16} =_q h_{13},\ h_{19} =_q h_{15},\ h_{25} =_q h_{23} \,\} \,\rangle
\end{array}
$$

[10] Note that while Turkish this is not an example of high nominalization, all three analyses presented in this section produce the same semantic representation.

This semantic structure contains the predication nominalization_rel and the verb is the first argument of this predication.[11] The intrinsic argument (ARG0) of the nominalization_rel is of type *x* for individual, rather than *e* for event, so as to be a suitable argument for adjectival modifiers and bound variable for quantifiers. Because the low and mid analysis allow for the attachment of adjectives and quantifiers syntactically, this must be accounted for in the semantics as well.

However, we provide the user with analytical freedom regarding the semantic structure, by developing an option for nominalization that is purely syntactic. In this case the unary rule changes the HEAD value to noun and creates a direct semantic identity between the mother and daughter without adding nominalization_rel, resulting in MRSs like the one shown in (21).[12]

(21)
$$\langle\, h_1, e_2,$$
$$\left|\begin{array}{l} h_3{:}\mathsf{pron\_rel}(\textsc{arg0}\ x_4), \\ h_5{:}\mathsf{exist\_q\_rel}(\textsc{arg0}\ x_4,\ \textsc{rstr}\ h_6,\ \textsc{body}\ h_7), \\ h_8{:}\mathsf{\_cinema\_n\_rel}(\textsc{arg0}\ x_9), \\ h_{10}{:}\mathsf{exist\_q\_rel}(\textsc{arg0}\ x_9,\ \textsc{rstr}\ h_{11},\ \textsc{body}\ h_{12}), \\ h_{13}{:}\mathsf{come\_v\_rel}(\textsc{arg0}\ e_1,\ \textsc{arg1}\ x_4,\ \textsc{arg2}\ x_9), \\ h_{23}{:}\mathsf{pron\_rel}(\textsc{arg0}\ x_{22}), \\ h_{24}{:}\mathsf{exist\_q\_rel}(\textsc{arg0}\ x_{22},\ \textsc{rstr}\ h_{25},\ \textsc{body}\ h_{26}), \\ h_{21}{:}\mathsf{want\_v\_rel}(\textsc{arg0}\ e_2,\ \textsc{arg1}\ x_{22},\ \textsc{arg2}\ h_{27}) \end{array}\right|$$
$$\{\, h_6 =_q h_3,\ h_{11} =_q h_8,\ h_{16} =_q h_{13},\ h_{25} =_q h_{23},\ h_{27} =_q h_{13} \,\} \,\rangle$$

4.7 Summary

This section has presented our cross-linguistic analysis of nominalization. As is typical for Grammar Matrix libraries, the analysis encompasses a range of options. These options accommodate both cross-linguistic variation in the underlying phenomenon and analytic variation, facilitating the exploration of different analyses within implemented grammars.

5 Implementation in the Grammar Matrix

We implemented the analyses described in §4 in the Grammar Matrix, such that the user can define multiple nominalization strategies that can be accessed by the subordinate clause libraries, including Clausal Complements (Zamaraeva et al., to appear) and Clausal Modifiers (Howell & Zamaraeva, 2018). The user can give each nominalization strategy a name and select the level and desired semantic representation for that strategy. This strategy can then be associated with morphological rules (corresponding to nominalization affixes) and clausal complement and

[11] This relationship is mediated by a so-called *qeq* constraint. See note 6.

[12] As Turkish does not in fact have high nominalization, this MRS would not be produced for Turkish. We provide it here for comparison with the one in (20) only.

clausal modifier strategies that require nominalization. The relevant portion of the Grammar Matrix web questionnaire is illustrated by figure 7.

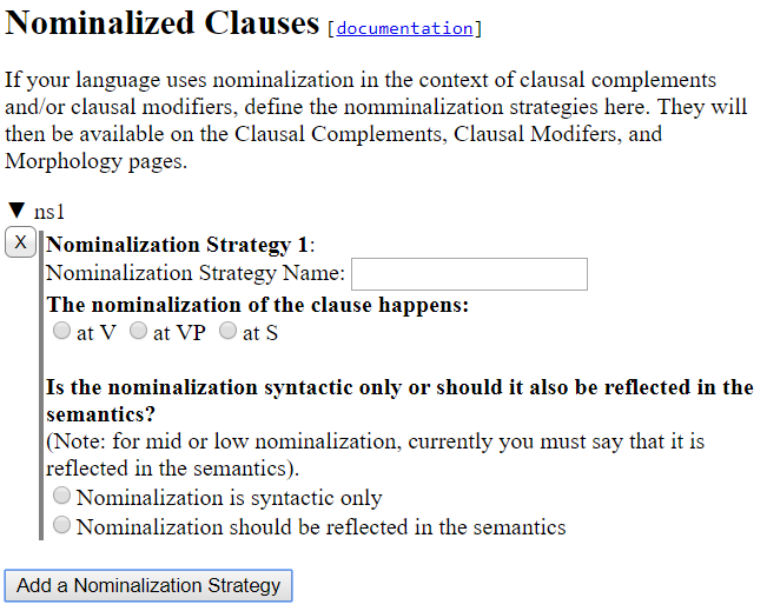


Figure 7: Snippet of Grammar Matrix questionnaire for nominalization library

6 Testing, Evaluation, and Error Analysis

Following typical practice in the development of Grammar Matrix libraries (Bender et al., 2010), we evaluated our implementation of this analysis by creating grammar fragments for a number of languages. This allows us to verify both that the analyses generalize to languages we didn't directly consider during library development and that the analyses in the library interact appropriately with other libraries.

We do initial verification using both artificial 'pseudolanguages' designed to test each combination of nominalization level and semantic representation and real languages. In both cases, we first develop testsuites including grammatical and ungrammatical examples, and then create choices files describing those languages. We feed the choices files to the Grammar Matrix customization system and use the resulting grammars to parse the testsuites using the LKB software (Copestake, 2002). Undergeneration, overgeneration, spurious ambiguity, or incorrect parses of testsuite items will indicate errors in the analysis or its implementation, which we fix during the development process.

We developed psudolanguage choices files and testsuites for each level of nominalization and each semantic representation for both nominalized clausal complements and clausal modifiers, resulting in a total of 8 pseudolanguages. For our

real language verification tests, we used Russian [rus] and Turkish [tur] for clausal complements, and Rukai [dru] for clausal modifiers. We refined our implementation until we achieved full coverage (all grammatical sentences correctly parsed) and no overgeneration (no ungrammatical sentences parsed) over the development testsuites. While the 8 pseudolanguage testsuites were targeted at nominalization, the real language testsuites contained examples for clausal complements or clausal modifiers in general, so not all examples were relevant to nominalization. The following table identifies both the overall results and those relevant specifically to nominalization.[13]

	Total		**Nominalized Clause**	
Language	**Coverage**	**Overgen.**	**Coverage**	**Overgen.**
Russian [rus]	6/6	0/11	6/6	0/11
Turkish [tur]	7/7	0/9	6/6	0/8
Rukai [dru]	2/2	8/8	2/2	8/8

Table 1: Results for development languages[14]

Finally, we tested our analysis on languages that we had not previously considered in order to evaluate how well it generalizes cross-linguistically. We consider evaluation to be extrinsic as it was evaluated as part of our evaluation for clausal complements (Zamaraeva et al., to appear) and clausal modifiers (Howell & Zamaraeva, 2018).[15] We evaluated our analysis in complement clauses in Yakima Sahaptin [yak] and Paresi-Haliti [pab], as well as in clausal modifiers in Basque [eus]. The results are presented in Table 2, again differentiating between the total number of examples and just those relevant to nominalization.

	Total		**Nominalized Clause**	
Language	**Coverage**	**Overgen.**	**Coverage**	**Overgen.**
Paresi-Haliti [pab]	5/5	0/6	3/3	0/4
Yakima Sahaptin [yak]	10/10	0/6	10/10	0/6
Basque [eus]	13/16	0/10	5/8	0/3

Table 2: Results for held-out languages[16]

The error analysis revealed one error (affecting three sentences in the testsuite for Basque), which was not directly related to the analysis presented in this

[13] We define "relevant" here as examples either containing a nominalized verb, or negative examples that are ungrammatical because they lack a nominalized verb.

[14] Russian, Turkic and Rukai are from the Indo-European, Altaic and Austronesian language families, respectively.

[15] More detailed discussion of the evaluation for those libraries beyond that which is relevant to nominalized clauses can be found in their respective papers.

[16] Paresi-Halit, Yakima Sahaptin and Basque are from the Arawaken, Penutian, and Basque language families, respectively.

paper, but revealed an interaction with another analysis stored in the Grammar Matrix. The Argument Optionality library (Saleem, 2010) adds phrase structure rules to grammar fragments that facilitate argument dropping. As this library was created before nominalized verbs were supported, these rules constrained the head-daughter to be [HEAD *verb*], thereby prohibiting subject dropping for nominalized verbs in Basque. We were able to confirm that these sentences would otherwise parse by adding a subject dropping rule to the grammar that allowed a nominalized verb to be the head daughter.

7 Conclusion

In this paper we present a cross-linguistic analysis of nominalization, designed to support analyses of this phenomenon as it appears in both clausal complements and clausal modifiers. The analysis is implemented in the form of a Grammar Matrix library and its interoperability with libraries for not just clausal complements and clausal modifiers but also other libraries including argument optionality, case, and word order is tested according to the standard Grammar Matrix evaluation methodology. We provided an analysis that allows nominalization to occur at three different levels in the syntax and provided two semantic representations. We plan to look at a wider range of languages as part of future work to determine the usefulness of the high nominalization analysis. We are also considering extending the option to omit nominalization from the semantics to the mid and low analyses, if we find evidence to do so. Our evaluation so far suggests that our analysis provides sufficient flexibility to handle both the typologically attested range of variation in this phenomenon and to provide a degree of analytical freedom to the linguist, while still maintaining comparability across language types.

Acknowledgments

We thank Edith Aldridge and the audience at HPSG 2018 for helpful comments and discussion. This material is based upon work supported by the National Science Foundation under Grant No. BCS-1561833. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Bender, Emily M., Scott Drellishak, Antske Fokkens, Laurie Poulson & Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation* 8. 1–50.

Bender, Emily M., Dan Flickinger & Stephan Oepen. 2002. The Grammar Matrix: An open-source starter-kit for the rapid development of cross-linguistically con-

sistent broad-coverage precision grammars. In John Carroll, Nelleke Oostdijk & Richard Sutcliffe (eds.), *Proceedings of the workshop on grammar engineering and evaluation at the 19th International Conference on Computational Linguistics*, 8–14. Taipei, Taiwan.

Bresnan, Joan. 1997. Mixed categories as head sharing constructions. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG'97 Conference*, Stanford: CSLI Publications.

Copestake, Ann. 2002. Definitions of typed feature structures. In Stephan Oepen, Dan Flickinger, Jun-ichi Tsujii & Hans Uszkoreit (eds.), *Collaborative language engineering*, 227–230. Stanford, CA: CSLI Publications.

Copestake, Ann, Dan Flickinger, Carl Pollard & Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation* 3(2-3). 281–332.

Howell, Kristen & Olga Zamaraeva. 2018. Clausal modifiers in the Grammar Matrix. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2939–2952.

Kornfilt, Jaklin. 1997. *Turkish*. London: Routledge.

Lapointe, Steven. 1993. Dual lexical categories and the syntax of mixed category phrases. In *Proceedings of the 10th Eastern States Conference of Linguistics*, 199–210. CLC Publications Ithaca, NY.

Malouf, Robert P. 1998. *Mixed categories in the hierarchical lexicon*. Palo Alto, CA: Stanford University dissertation.

Noonan, Michael. 2007. Complementation. In Timothy Shopen (ed.), *Language typology and syntactic description. volume 2: Complex constructions*, 52–150. Cambridge University Press.

Pollard, Carl & Ivan A Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

Pullum, Geoffrey K. 1991. English nominal gerund phrases as noun phrases with verb-phrase heads. *Linguistics* 29(5). 763–800.

Saleem, Safiyyah. 2010. *Argument optionality: A new library for the Grammar Matrix customization system*. University of Washington MA thesis.

Zamaraeva, Olga, Kristen Howell & Emily M. Bender. to appear. Modeling clausal complementation for a grammar engineering resource. To appear in *Proceedings of the 2nd meeting of the Society for Computation in Linguistics*.

Zeitoun, Elizabeth. 2007. *A grammar of Mantauran (Rukai)*. Institute of Linguistics, Academia Sinica Taipei.