# Dependency annotation of noun incorporation in polysynthetic languages

**Francis M. Tyers**
Department of Linguistics
Indiana University
Bloomington, IN
ftyers@iu.edu

**Karina Mishchenkova**
School of Linguistics
Higher School of Economics
Moscow
karinam6@mail.ru

## Abstract

This paper describes an approach to annotating noun incorporation in Universal Dependencies. It motivates the need to annotate this particular morphosyntactic phenomenon and justifies it with respect to frequency of the construction. A case study is presented in which the proposed annotation scheme is applied to a corpus of Chukchi, a highly-endangered language of Siberia that exhibits noun incorporation. We compare argument encoding in Chukchi, English and Russian and find that while in English and Russian discourse elements are primarily tracked through noun phrases and pronouns, in Chukchi they are tracked through agreement marking and incorporation, with a lesser role for noun phrases.

## 1 Introduction

This paper addresses the question of noun incorportion in Universal Dependencies. It gives an overview of the phenomenon and some current challenges with its representation in Universal Dependencies. It then describes an annotation solution requiring minimal changes to the existing annotation guidelines. A case study is then given in which a corpus of an endangered polysynthetic language that exhibits noun incorporation is annotated and a comparison is drawn between how this language encodes arguments and how two more well-studied and better-resourced languages do.

There are many definitions of polysynthetic languages, and there is no agreement in the literature as to what precise features of a language merit its inclusion within the category of polysynthetic languages (Fortescue et al., 2017). A common definition is a language is polysynthetic if its verbal morphology is extremely complex and a single verb is capable of expressing all of its arguments internally, through agreement or incorporation, i.e. *holophrasis*.

In comparison to more familiar morphological types such as isolating, fusional and agglutinative languages, vanishingly little computational work has been done on languages of this type, although note the recent workshop on polysynthetic languages (Klavans, 2018). This is largely as a result of the fact that these languages are usually spoken by communities that are either comprised of a small number of speakers, have limited economic and political power or both. Geographically, there are few if any in Europe or most of Asia, but they are spoken by many indigenous communities in the Arctic region, the Americas, Australia and Papua New Guinea (Fortescue et al., 2017).

We present the first work on annotating noun incorporation in a language[1] using the Universal Dependencies guidelines — and probably the first medium-scale computational annotation work of any kind of a language of the noun-incorporating type.[2]



[1] Senuma and Aizawa (2017) present a treebank of Ainu, a language that is recorded as exhibiting noun incorporation. The treebank contains 36 sentences. The authors state that noun incorporation for Ainu is only used in poetry and fixed expressions but provide no quantative evidence. Note that this existence of incorporation as a marginal phenomenon is not the case for all languages that exhibit incorporation.

[2] We note that Bick (2019) presents a dependency formalism for Greenlandic, which has a phenomenon similar to noun incorporation, but that we distinguish as lexical affixing — a large, but closed, set of verbalising affixes that can be attached to nouns to make complex predicates.



*Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 194–203

In passing we address two issues that are often brought up when discussing the phenomenon of noun incorporation. The first is a matter of frequency: is this a core part of the language, or is it a marginal phenomenon? This matters because if we are to propose changing annotation guidelines that may have an effect on hundreds of existing treebanks, then our approach will be more convincing if the phenomenon is frequent and the change is minimal. The second is a matter of level of representation: is this syntax, morphology or something else? This matters because (typically) if it is grouped within the syntactic phenomena then it should have an expression in the syntactic annotation, but if it is grouped with morphology then it should be expressed in the morphological annotation.

## 2 Noun incorporation

Noun incorporation is a phenomenon whereby a noun and a verb are combined to produce a complex verb. It forms part of a wider group of incorporation phenomena such as nominals being incorporated with other nominals, a process often called *compounding*. In this paper we are specifically concerned with the incorporation of core arguments into verbal predicates. Or put another way, the 'saturation' of argument slots in the verb by compounded lexical material.

| | | | |
|---|---|---|---|
| | Ынӷэната | гакинораԓенатъым | кԓюпчыко. |
| | *ənqenata* | *ɣakinoraԓenatʔəm* | *kԓupsəko* |
| (1) | *ənqena-ta* | *ɣa-kino-rat-ԓena-t=ʔəm* | *kԓup-səko* |
| | this-INS | PF-film-bring-PF.3SG-PL=EMPH | club-LOC |
| | 'They brought the film to the club by this transport.'[3] | | |

We can illustrate this process with Example 1 from Chukchi, the form гакинораԓенатъым [ɣakinoraԓenatʔəm] is composed of two lexical roots, *кино* 'film' and *-рэт-* [ret] 'bring'.[4] The process of incorporation in this case renders a transitive verb, intransitive, with the concomitant effects on agreement inflection — Chukchi has separate inflectional paradigms for transitive and intransitive verbs. Here, the circumfix for third-person plural subject in the perfect aspect of the stative paradigm is *гэ- … -ԓиӈэт* [ɣe- … -ɬine-t].

There is a question in the literature as to what extent the incorporated noun can have definite reference, which overlaps with the discussion of if noun incorporation should be considered primarily a syntactic phenomenon or a lexical phenomenon. There has been a substantial amount of lively debate in the literature written about this with authors such as Baker (1996) positing that noun incorporation is a syntactic movement rule which takes a direct object and moves it inside the verbal complex leaving a trace. In this analysis, the incorporation of nouns with definite reference is permitted, and in some cases obligatory.

Dunn (1999) in his description of Chukchi draws a distinction between lexical incorporation, or compounding (see above) and *syntactic incorporation*, noting that syntactic incorporation leads to a rearrangement of valency in the verb and that there can be "distinguished dependency relationships between the two stems" (Dunn, 1999, §12.1).

On the other end of the scale, incorporation has been treated as a lexical phenomenon (Mithun, 1984; Rosen, 1989; Anderson, 2001), although they also acknowledge the discourse and pragmatic usage of noun incorporation and the part it plays in the argument structure of predicates. Mithun (1984) in particular provides a four-way categorisation of noun incorporation ranging from the more lexical to the more syntactic. Different languages may exhibit different kinds of incorporation. In this paper we are primarily concerned with Mithun's 'TYPE III' incorporation, or that incorporation which is used to manipulate the discourse structure, with the incorporated noun having low discourse salience.

Noting that discussion of the issue has been caught up in various debates surrounding theoretical syntax, we prefer to take a practical approach. Noun incorporation is a wide-ranging phenomenon with effects in morphology — the form of the word, syntax — the arrangement of clause and argument structure, and discourse — the arrangement of information structure. Thus we find a purely lexical approach, or the 'all verbs with incorporated nouns should be annotated as separate lexemes' to be untenable.

[3] INS – instrumental; PF – perfective; 3SG – third person singular; PL – plural; EMPH – emphasiser; LOC – locative. Hyphens denote morpheme boundaries, while the equals sign denotes a clitic boundary.

[4] The transformation of the vowel in the verb stem *-рэт-* [ret] 'bring' to *-рат-* [rat] is a vowel harmony process. For a short description of vowel harmony in Chukchi, refer to §4.

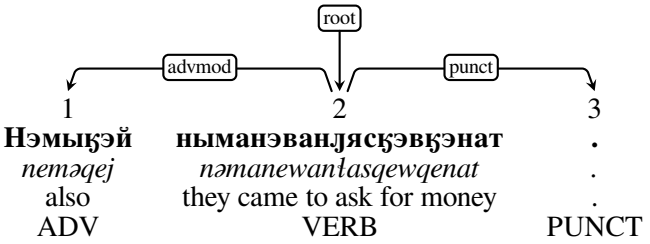


**Figure 1:** A simple dependency annotation scheme. The transitive verb *-ваԯ-* [wanɬa] 'ask' has been combined with the lexical stem *маӈэ* [mane] 'money' to produce a new intransitive verb, to which is added intransitive agreement morphology.

## 3 Proposed annotation scheme

In order to develop and test annotation guidelines for these phenomena, we decided to approach a particular language, Chukchi (see Section 4), and develop them iteratively during an annotation project. For the base annotation guidelines we used those of the Universal Dependencies project (Nivre et al., 2020) and extended them following the six principles of *Manning's Law*:

1. UD needs to be satisfactory for analysis of individual languages.
2. UD needs to be good for linguistic typology.
3. UD must be suitable for rapid, consistent annotation.
4. UD must be suitable for computer parsing with high accuracy.
5. UD must be easily comprehended and used by a non-linguist.
6. UD must provide good support for downstream NLP tasks.

Within the current guidelines for Universal Dependencies, a suggested approach for annotating noun incorporation is of the strong lexicalist type. Relations are between syntactic *words*. For a language exhibiting noun incorporation this would result in trees such as in Figure 1.

Here, 'to come to ask for money' would be represented a single verb. We argue that this is not a satisfactory analysis (1), and that it is not useful for downstream NLP tasks (6). As a result of the lack of information in the annotation, it would be suitable for rapid, consistent annotation (3) and high-accuracy parsing (4). While in terms of comprehensibility for non-linguists (5) and use for linguistic typology (2) various arguments may be made, there is not much to understand in the annotation, and it may be useful in terms of illustrating that certain languages have very long verbs, but not in terms of looking at anything more than morphology from a typological point of view.

It is worth presenting for a moment the lexicalist hypothesis to which Universal Dependencies subscribes. Broadly stated it is that syntactic structures or relations hold between *words*, that is there is a separate component (in the human mind) for building words — the *lexicon* — and for building sentences — the *grammar*. The strong variant of this hypothesis states that all production of word forms happens in the lexicon, and that this contains lexemes, derivational rules,[5] and inflectional rules. Syntactic rules may not interact with derivational rules. This is often described as the 'Lexical Integrity Principle' and is adopted by formalisms such as Head-Driven Phrase Structure Grammar (HPSG; Bresnan and Mchombo (1995)) (although see also (Emerson and Copestake, 2015)) and Lexical-Functional Grammar (LFG; Dalrymple (2001)). A weak variant of the lexicalist hypothesis states that lexemes and derivational rules belong in the lexicon, while syntax operates on structures composed of lexemes and features, to which subsequently are applied inflectional rules. Finally, there is non-lexicalism, in which syntax applies directly to simple lexemes and morphemes, a popular approach in contemporary theoretical syntax.

The Universal Dependencies project as a whole subscribes to strong lexicalism, that is syntactic structures and relations hold between syntactic words, although there is flexibility with how the notion of *word*[6] is defined. For example, clitics are often, but not always, tokenised as separate syntactic words.

[5]Here we refer to morphological derivation, these rules are sometimes called *word-formation rules*.

[6]We note here that there is discussion as to if the word should be a unit of analysis at all. For an informative overview, see Haspelmath (2009).

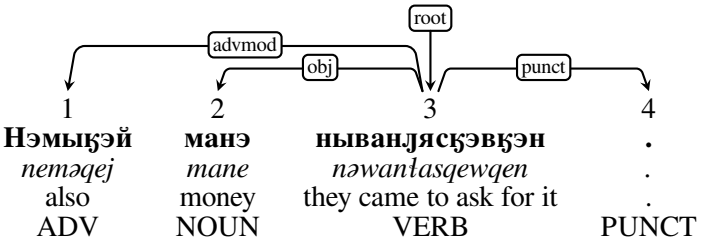


**Figure 2:** An annotation scheme where the intransitive clause post-incorporation has been rewritten as a transitive clause with no incorporation. Note that in addition to moving the incorporated element out of the verb, the agreement circumfix also must be modified. The intransitive stative habitual agreement circumfix for the third-person plural subject is *n*-STEM-*qinet*, whereas the transitive agreement for third-person plural subject and third-person singular object is *n*-STEM-*qin*. Thus, producing this annotation would necessitate the reinflection of the verb by the annotator.

```
# sent_id = Money:10
# text = Нэмыӷэй ныманэванԯясҷэвӷэнат.
# text[phon] = nemәqej nәmanewanɬasqewqenat
# text[rus] = Тоже приходили просить денег.
# text[eng] = They also came to ask for money.
1    Нэмыӷэй              _   ADV    _   _   4   advmod   _   _
2-4  ныманэванԯясҷэвӷэнат _   _      _   _   _   _        _   _
3    манэ                 _   NOUN   _   _   4   obj      _   _
4    ныванԯясҷэвӷэн       _   VERB   _   _   0   root     _   _
5    .                    _   PUNCT  _   _   4   punct    _   _
```

**Figure 3:** Partial CoNLL-U representation of the tree in Figure 2 illustrating the encoding of the verb as a multi-token syntactic word. The contents of the FEATS column has been omitted for reasons of space.

If we accept that noun incorporation should be encoded in the annotation, the question then becomes, where should it be encoded? A morphological encoding would see the incorporated noun become part of the morphological features, possibly with a expression like `Incorporated[obj]=манэ` for 'the object of this verb has been incorporated and the lexeme is *манэ*'. An advantage of this method would be that it is minimally disruptive to the existing guidelines, adding additional language-specific `Feature=Value` pairs is directly permitted by the guidelines. However, it has some unsatisfactory consequences, typically the `Value` portion of `Feature=Value` pairs are considered a finite set. There is a fixed number of possible `Values` for each `Feature`. Noun incorporation does not follow this. In languages that permit it, any noun — semantic and pragmatic conditions permitting — can be incorporated as an object.

A second option is to include it in the basic dependencies, using the existing solution for 'multiword' tokens, clitics, and contractions. This would involve splitting the token into sub-tokens and annotating as the underlying construction, for example an intransitive clause would be annotated as if it were transitive with a free-standing object. This is primarily unsatisfactory from a descriptive point of view, there is every indication, regardless of the theory subscribed to, that the surface structure of a transitive verb with incorporated object is intransitive.[7] Additionally, it would require substantial effort on the part of annotators, who would need to be both fluent in the language (to be able to generate the non-incorporated equivalents of clauses with incorporation), and expert in the annotation scheme. For many (if not most) under-resourced and marginalised languages, this does not obtain. Furthermore, the information structure of the incorporated and non-incorporated forms are different, so further processing would need to distinguish unconverted transitive clauses and converted ones.[8]

Our proposal is to encode it in the *enhanced* dependency structure (Schuster and Manning, 2016). The enhanced structure is built on top of the basic dependencies and may include:

[7]For simplicity we restrict discussion of antipassivising verbs (Dunn, 1999, §12.2.1) here, where the incorporation of an object does not reduce the valency, and some other non-patient role is promoted to the object slot.

[8]To aid the reader, one could imagine a hypothetical annotation scheme whereby analytic passives must be rewritten as non-passive with a feature indicating passivisation.

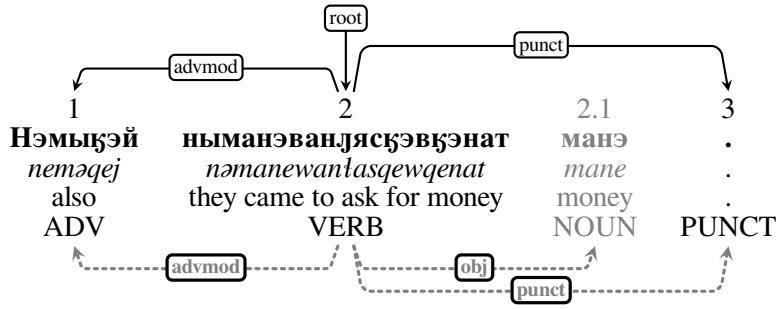


**Figure 4:** Dependency tree for the sentence in Figure 5. The enhanced representation is shown in grey. As for morphological features, the verb is marked with a feature `Incorporated[obj]=Yes` and a feature `Valency=1` to indicate the intransitive nature of the verb. The incorporated noun in the enhanced representation receives the feature `Incorporated=Yes`.

1. Null nodes for elided predicates
2. Propagation of conjuncts
3. Additional subject relations for control and raising constructions
4. Coreference in relative clause constructions
5. Modifier labels that contain the preposition or other case-marking information

We propose extending the guidelines for the enhanced representation to allow *additional* nodes for core arguments of predicates, which are expressed via incorporation of lexical material.[9] Note that these are not strictly *null* nodes — such as those used for elided predicates — as they could only be permitted to represent incorporated lexical material, which by its nature is not *null*. This would allow the annotation of trees such as that in Figure 4 where the incorporated object becomes a node in the enhanced graph.

In an outward sense, the annotation of incorporation has some relation to the annotation of *pro-drop* languages, where arguments required by the predicate may not have any form in the syntax and only appear as agreement markers on the verb. However in one important sense it differs in that while for pro-drop languages the potential list of pronouns is from a finite set and can often be inferred mechanically from the verbal agreement, with incorporation the arguments are not a finite set and, barring additional annotation, cannot be recovered from the predicate.

## 4 Case study

In order to test our proposed annotation guidelines, we decided to approach a particular language, Chukchi. Chukchi (ISO-639-3: `ckt`) is a highly endangered and polysynthetic language spoken in the sparsely-populated Chukotka Autonomous Okrug in the far north east of the Russian Federation. The total population of Chukotka was 50,526 in 2010. According to the 2010 census it was spoken by 5,095 people, or around a third of the ethnic population. Today most speakers are over the age of 50, and, even by the 1990s intergenerational transmission had been disrupted (Dunn, 1999). The language exhibits polypersonal agreement, ergative–absolutive alignment, and a subject–object–verb basic word order in transitive clauses. The language is severely under-resourced and there has been very little computational work on this language. We are only aware of a description of a finite-state morphological analyser (Andriyanets and Tyers, 2018). There have been a number of theoretical and descriptive linguistic works on noun incorporation in Chukchi, including Spencer (1995) who gives a general overview and Polinsky (1990) who covers subject incorporation.

We used the Amguema corpus, available through the «Chuklang»[10] site, which is a corpus of spoken Chukchi in the Amguema variant. The corpus consists of both audio recordings and transcriptions with glosses and translations in Russian and English. There are a total of 65 texts, most of which are elicited

[9] This is the most conservative variant of our proposal, the most essential part. We also think it is worth opening up a discussion about *null* nodes for core arguments expressed morphologically, such as subject and object in languages with polypersonal agreement.

[10] `https://chuklang.ru/`

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1.10 | neməqej | nəmanewanłasqewqenat | | | | | |
| | neməqej | nə- | mane | wanła | -sqew | -qena | -t |
| | тоже | ST | деньги | просить | MCP | ST.3SG | PL |
| | also | ST | money | ask | MCP | ST.3SG | PL |
| | ‘Тоже приходили просить денег.’ | | | | | | |
| | ‘They also came to ask for money.’ | | | | | | |

**Figure 5:** An annotated sentence in Chukchi from the Amguema corpus, text *Деньги* ‘Money’. The sentence includes an ID, a phonetic transcription, morpheme segmentation, gloss in Russian and English and a free translation in Russian and English. The sentence demonstrates object incorporation, the object *-mane-* ‘money’ is combined with the transitive stem *-wanła-* ‘ask’ to make an intransitive verb which is then conjugated with subject conjugation for 3rd person plural *nə- …-qena-t*.

stories and tales, comprising 1,004 sentences/utterances with 6,124 tokens. The corpus was created between 2016 and 2018 by Chukchi speakers and researchers from *Higher School of Economics* in Moscow.

Figure 5 presents an example of a sentence from one of the texts in the Amguema corpus. In this sentence, the noun *мане* [mane] ‘money’ has been incorporated as an object of the verb *-ванӆя-* [wanła] ‘ask’; the derivational affix, *-сқев* [sqew] ‘GOAL’, is suffixed, and the inflectional agreement morphology is circumfixed.

The tokenisation in the corpus follows the scheme set out by Dunn (1999) and others, in that the formal boundary of a word is indicated by the vowel harmony process. For an extensive description the reader is referred to Dunn (1999, §3.4.1), but in brief: In Chukchi vowels are split into two groups, recessive, *и* /i/, *э* /e$_1$/ and *ы* /u/ and dominant *э* /e$_2$/, *а* /a/ and *о* /o/. The two variants of /e/ are phonetically identical but phonologically behave differently. If any vowel in any morpheme in a word is dominant, then any recessive vowels harmonise to their dominant counterparts.

Nouns incorporated into verbs, as with all other morphemes in the verb form, participate in this process. Consider the example *тыӈырэӆқытатқэнат* [taŋnəreɬqəpatqenat] ‘They cooked porridge’. The incorporated nominal object *-рэӆқ-* [-reɬq-] < *риӆық* [riɬəq] ‘porridge’, which is harmonically recessive, undergoes *и* /i/ → *э* /e/ harmony as a result of the dominant vowel /a/ in the verb stem *-нат-* [pat] ‘cook’.

The corpus was annotated for dependency structure by two linguists over a period of around two months. Each linguist took a disjunct set of texts to annotate. After the dependency structure was annotated, a program was written to convert the glosses into parts of speech and sets of morphological features.

There were a total of 79 incorporated elements in the corpus, which leads to a per token percentage of 1.2%, and a per utterance percentage of 7.8%. If we look at the percentage of verb forms with incorporated elements, the percentage is 6.6%.[11] Around half of all texts contained no incorporations, and around half contained more than one with the minimum being 0 and the maximum being 14. We aim to show with these statistics that although the per token percentage may appear to be marginal, if we look at the level of predicates and discourse, the phenomenon is far from marginal and is a core part of the language.

By and far the most productive type of incorporation was object incorporation, with 50 out of 79 examples. Following this was incorporation of verb stems as adverbial modifiers, which we do not treat here. More marginal, under five examples each were incorporation of obliques, subjects and adverbs.

Figure 4 presents a CoNLL-U representation of the tree in Figure 4. Lemmas have yet to be included, and the morphological features are `Aspect=Hab` ‘Habitual aspect’, `Deriv[goal]=Yes` ‘Goal derivation’, `Incorporated[obj]=Yes` ‘Incorporated object’, `Mood=Ind` ‘Indicative mood’, `Number[subj]=Plur` ‘Plural subject’, `Person[subj]=3` ‘Third-person subject’, `Valency=1` ‘Intransitive’, `VerbForm=Fin` ‘Finite verb form’, `Voice=Stat` ‘Stative verbal paradigm’. The goal derivation indicates motation towards a goal.

## 5 Comparison

To illustrate some differences between how Chukchi encodes arguments and how English and Russian do, we selected a short story from the corpus and categorised how different entities (principally subjects and objects)

[11] We note that this percentage far exceeds that than phenomena such as reflexive pronouns in English, which account for under 1% of all pronouns, but without which an annotation scheme for English could hardly be considered complete.

```
# sent_id = Money:10
# text = Нэмыӄэй ныманэванӆясӄэвӄэнат.
# text[phon] = neməqej nəmanewanɬasqewqenat
# text[rus] = Тоже приходили просить денег.
# text[eng] = They also came to ask for money.
1    Нэмыӄэй              _   ADV    _   _    2   advmod   2:advmod  _
2    ныманэванӆясӄэвӄэнат _   VERB   _   _    0   root     0:root    _
2.1  манэ                 _   NOUN   _   _    _   _        2:obj     _
3    .                    _   PUNCT  _   _    2   punct    2:punct   _
```

**Figure 6:** Partial CoNLL-U representation of the tree in Figure 4 illustrating the encoding of the incorporated object *манэ* 'money' in the enhanced representation. The contents of the FEATS column has been omitted for reasons of space. The `2.1` notation is usually used for elided predicates, but here we extend it to incorporated objects.

are encoded. These were split into four categories: Non-incorporated nominals, incorporated nominals, agreement affixes and pronominals.

The annotation is shown in Figure 7. The motivation behind this comparison is to demonstrate in a visually interpretable way the necessity of an annotation scheme that includes information about incorporated nouns.

In this comparison we can clearly see that English and Russian both have strong tendencies towards encoding arguments with free pronouns. The majority of sentences have at least one pronominal argument. In addition, Russian makes use of verbal agreement markers to encode the subject. English also uses agreement markers, but sparingly. Most verb forms are not inflected for person and number.

In Chukchi however, fewer than half of all sentences contain an explicit pronoun or external noun phrase argument. Arguments are either encoded via incorporation, agreement or by a combination of these two processes. As Dunn (1999, §7.2) observes, personal pronouns are "textually rare and pragmatically marked", and "in unelicited texts [...] are not used for anaphoric specification of arguments in clauses".

This clearly has implications for language technology applications and further linguistic analysis. Annotation schemes for predicate–argument structure such as PropBank (Palmer et al., 2005; Haverinen et al., 2015) are often annotated over the tree structure, and systems for co-reference resolution such as *Xrenner* (Zeldes and Zhang, 2016) rely on the dependency structure to add co-reference information. Semantic parsing systems such as *Universal Semantic Parsing* (Reddy et al., 2017) also rely on the dependency structure. If incorporated objects are kept out of the tree structure, then specific language-specific solutions will have to be made in each downstream application for the languages that exhibit incorporation. It is our belief that the most adequate place to represent this information is in the morphosyntactic structure as part of the dependency tree.

## 6 Future work

We have been able to obtain permission to annotate a further corpus of Chukchi. This corpus contains an additional 1,000 sentences (approx. 11,500 tokens) of parallel text in Chukchi and Russian from the Chukotkan newspaper *Крайний Север* 'Krayny j Sever'.[12] It is reported by Comrie (1981) that incorporation in Chukchi is on the wane, and by Dunn (1999) that this may be the case for written Chukchi, but not spoken Chukchi, there has to our knowledge been no empirical study of this.

An additional aspect of future work is how to represent lemmas. The problem being that if we use the lemma from the basic representation then it should include the incorporated item, but then in the enhanced representation the object will be doubled: once in the lemma and once in the tree. However if we use the base lemma, then in the basic representation the lemma will not match the word form. Ideally for the verbal predicate we would like to have a different lemma in the enhanced representation to the basic representation. This would leave the basic representation with the lemma of the verb + incorporated element, and the enhanced representation with the root lemma of the verbal predicate.

Finally, in many polysynthetic languages, including Chukchi, there is a related process which also makes

[12] `https://www.ks87.ru/`

| | Language | Argument(s) | | | |
|---|---|---|---|---|---|
| | **Chukchi:** | Nom | Inc | Agr | Pro |
| 1 | Ӷонпы нывичвэтчыӷивӷинэтъым ныманэванӆясӷэвӷэнат. | - | + | ++ | - |
| 2 | Ӷынвэт [Ирана] нинэӆьуӷин [чиниткин экык] эймэвыӆьын ыныкагтым. | ++ | - | + | + |
| 3 | Нъыръыйыӆӷыӈӈоӷэн ынӈин вай. | - | - | + | - |
| 4 | [Эккэтэм] ынӈин ынӷэн нинэгитэӷинъым. | + | - | + | - |
| 5 | Ӷынвэт ӷынвэт нытанӈыткоӈӈоӷэн. | - | - | + | - |
| 6 | «[Ыммэмы] ӷэнаманэӆпынрыгэ.» | + | + | + | - |
| 7 | «Ынкы ӷээӷынъым мывичвысӷиквъэк автомат.» | - | - | + | - |
| 8 | Эээ ӷэӆюӷъым нэнаманэӆпынрыӷэн. | - | + | + | - |
| 9 | [Мургинэт] нэмыӷэй ныетӷинэт [ӈинӷэгти]. | + | - | + | - |
| 10 | Нэмыӷэй ныманэванӆясӷэвӷэнат. | - | + | + | - |
| 11 | «Э’тки вай ӷээӷын мытыӆгирывичвэнӈыркын.» | - | - | + | - |
| 12 | Ну ӷэӆюӷъым [ытӆыгэ] нэнаманэӆпынрыӷэнатэ амъянра наӷам. | - | + | + | - |
| 13 | Нэмэ ынӷэн комната ныйъоӷэн. | - | + | - | |
| | **Russian:** | Nom | Inc | Agr | Pro |
| 1 | Они постоянно ходили играть, постоянно просили [денег]. | + | - | ++ | + |
| 2 | И тут [Ира] видит, как [её сын] приближается к ней. | ++ | - | + + | + |
| 3 | Она стала делать вид, что она спит вот так. | - | - | +* | ++ |
| 4 | [Сын] стал вот так разглядывать её. | + | - | + | + |
| 5 | Наконец, она начала смеяться. | - | - | + | + |
| 6 | «[Мама], дай мне [денег].» | ++ | - | - | + |
| 7 | «Я пойду ещё там поиграю.» | - | - | ++ | + |
| 8 | Конечно, она ему дала [денег]. | + | - | + | + |
| 9 | [Наши сыновья] тоже подходили. | + | - | + | - |
| 10 | Тоже просили [денег]. | + | - | + | - |
| 11 | «Мы ужас как хотим поиграть ещё.» | - | - | + | + |
| 12 | Ну конечно, [отец] дал [денег], причём каждому отдельно. | ++ | - | + | - |
| 13 | И опять [эта тройка] направляется к комнате. | + | - | - | - |
| | **English:** | Nom | Inc | Agr | Pro |
| 1 | They constantly went to play, constantly asked for [money]. | + | - | - | + |
| 2 | And then [Ira] sees [her son] approaching her. | ++ | - | + | + |
| 3 | [She] began to pretend that she was sleeping like this. | - | - | + | ++ |
| 4 | [Her son] began to look at her like this. | + | - | - | + |
| 5 | Finally, she began to laugh. | - | - | - | + |
| 6 | “[Mum], give me [some money].” | ++ | - | - | + |
| 7 | “I’ll go and play there again.” | - | - | - | + |
| 8 | Of course, she gave him [the money]. | + | - | - | ++ |
| 9 | [Our sons] also came. | + | - | - | - |
| 10 | They asked for [money]. | + | - | - | + |
| 11 | “We really want to play more.” | - | - | - | + |
| 12 | Well, of course, [their father] gave them [money], separately. | ++ | - | - | + |
| 13 | And once again [these three boys] are heading to the room. | + | - | + | - |

**Figure 7:** Argument encoding in a short story: A comparison between encoding strategies in Chukchi, Russian and English. Entities are colour coded, inflectional agreement markers are underlined, noun phrases are in square brackets **[…]** and pronouns are unadorned. False starts have been removed from the Chukchi example. The columns on the right show four strategies for argument encoding, as a (Nom)inal, as (Inc)orporation, as (Agr)eement and as (Pro)nominal. In Chukchi, answering a question like “What did their sons want to do with the money?” would require a model of both inflection and incorporation, whereas in Russian and English the question could be answered on the level of tokens. Note that in (9) in Chukchi, *[Мургинэт …ӈинӷэгти]* ‘our …son’ is a discontinuous constituent. The glossing of this story can be found in `https://chuklang.ru/media/texts/Money.pdf` and the Latin transcription of Chukchi can be found in Appendix A.

compound predicates, lexical affixing. In this process a grammaticalised set of verbalising lexical affixes can be added to nouns to make verbs where the noun fills one of the valency slots. This is widely used in Chukchi and in other polysynthetic languages, such as Greenlandic and Yupik and has been treated before in formalisms such as HPSG (Malouf, 1999) and LFG (Grimshaw and Mester, 1985). The challenge with this construction is which verb to consider the head, as unlike with canonical noun incorporation it is not clear. The morphological head is certainly the root (the affixes are bound morphemes), while the semantic head is the verbalising affix (supplying the argument structure).

In addition to continuing work on Chukchi, we plan to work with other languages which exhibit incorporation, such as Western Sierra Nahuatl and Mapudungun.

## 7 Concluding remarks

In this paper we have presented an approach to dependency annotation of the phenomenon of noun incorporation within the Universal Dependencies framework. The approach modifies the existing guidelines by allowing core arguments expressed by noun incorporation to be included as additional nodes in the enhanced representation. Additionally we perform a case study using our annotation scheme by annotating the Amguema corpus of Chukchi and show that noun incorporation is not a marginal phenomenon.

## Acknowledgements

We would like to thank the Universal Dependencies community for very informative and useful discussions. We also thank the anonymous reviewers, Robert Pugh, and Kevin Scannell for their helpful and insightful comments. This article contains output of a research project implemented as part of the Basic Research Programme at the National Research University Higher School of Economics (HSE University).

## References

Stephen R. Anderson. 2001. Lexicalism, incorporated (or incorporation, lexicalized). In *Proceedings of the 36th Annual Meeting of the Chicago Linguistics Society*, pages 13–34.

Vasilisa Andriyanets and Francis M. Tyers. 2018. A prototype finite-state morphological analyser for Chukchi. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 31–40, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Mark Baker. 1996. *The Polysynthesis Parameter*. Oxford University Press.

Eckhard Bick. 2019. Dependency trees for Greenlandic. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 140–148.

Joan Bresnan and Sam A. Mchombo. 1995. The lexical integrity principle: Evidence from Bantu. *Natural Language and Linguistic Theory*, 13(2):181–254.

Bernard Comrie. 1981. *Languages of the Soviet Union*. Cambridge University Press.

Mary Dalrymple. 2001. *Lexical Functional Grammar*. Academic Press.

Michael Dunn. 1999. *A Grammar of Chukchi*. Ph.D. thesis, Australian National University.

Guy Emerson and Ann Copestake. 2015. Lacking integrity: HPSG as a morphosyntactic theory. In Stefan Müller, editor, *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar*, pages 75–95. CSLI Publications.

Michael Fortescue, Marianne Mithun, and Nicholas Evans. 2017. The Oxford Handbook of Polysynthesis.

Jane Grimshaw and Ralf-Armin Mester. 1985. Complex verb formation in Eskimo. *Natural Language and Linguistic Theory*, 3:1–19.

Martin Haspelmath. 2009. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1).

K. Haverinen, J. Kanerva, S. Kohonen, A. Missilä, S. Ojala, T. Viljanen, V. Laippala, and F. Ginter. 2015. The Finnish Proposition Bank. *Language Resources and Evaluation*, 49:907–926.

| | Language | Argument(s) | | | |
|---|---|---|---|---|---|
| | **Chukchi:** | Nom | Inc | Agr | Pro |
| 1 | Qonpə nəwiswetsəqiwqinetʔəm nəmanewanłasqewqenat. | - | + | ++ | - |
| 2 | Qənwet [Irana] inełʔuqin [sinitkin ekək] ejmewəłʔən ənəkaɣtəʔm. | ++ | - | + | + |
| 3 | Nʔərʔəjəłqəŋŋoqen ənŋin waj. | - | - | + | - |
| 4 | [Ekketeʔm] ənŋin ənqen ineɣiteqinʔəm. | + | - | + | - |
| 5 | Qənwet qənwet nətannŋətkoŋŋoqen. | - | - | + | - |
| 6 | «[Əmmemə], qenamanełpənrəɣe.» | + | + | + | - |
| 7 | «Ənkə qeeqənʔəm məwiswəsqikwʔek avtomat.» | - | - | + | - |
| 8 | Eee qełuqʔəm nenamanełpənrəqen. | - | + | + | - |
| 9 | [Muryinet] neməqej nəjetqinet [ŋinqeɣti]. | + | - | + | - |
| 10 | Neməqej nəmanewanłasqewqenat. | - | + | + | - |
| 11 | «ʔetki waj qeeqən mətəłɣirəwiswenŋərkən.» | - | - | + | - |
| 12 | Nu qełuqʔəm [ətłəɣe] nenamanełpənrəqenate amʔjanra naqam. | - | + | + | - |
| 13 | Neme ənqen komnata nəjʔoqen. | - | - | + | - |

**Figure 8:** Argument encoding in a short story: Chukchi have been orthographically transliterated into Latin script.


Judith L. Klavans, editor. 2018. *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Robert Malouf. 1999. West Greenlandic noun incorporation in a monohierarchical theory of grammar. In Gert Webelhuth, Andreas Kathol, and Jean-Pierre Koenig, editors, *Lexical and Constructional Aspects of Linguistic Explanation*, pages 47–62. CSLI Publications.

Marianne Mithun. 1984. The evolution of noun incorporation. *Language*, 60(4):847–894.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Chris Manning, Sampo Pyysalo, Sebastian Schuster, Francis M. Tyers, and Dan Zeman. 2020. Universal Dependencies v2: an evergrowing multilingual treebank collection. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4027–4036.

Martha Palmer, P. Kingsbury, and D. Gildea. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Maria Polinsky. 1990. Subject incorporation: Evidence from Chukchee. In Katarzyna Dziwirek, Patrick Farrell, and Errapel Mejias-Bikandi, editors, *Grammatical relations: A cross-theoretical perspective*, pages 349–364.

Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. Universal Semantic Parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 89–101, Copenhagen, Denmark, September. Association for Computational Linguistics.

Sara Thomas Rosen. 1989. Two types of noun incorporation: A lexical analysis. *Language*, 64:294–317.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Hajime Senuma and Akiko Aizawa. 2017. Toward Universal Dependencies for Ainu. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 133–139, Gothenburg, Sweden, May. Association for Computational Linguistics.

Andrew Spencer. 1995. Incorporation in Chukchi. *Language*, 71(3):439–489.

Amir Zeldes and Shuo Zhang. 2016. When annotation schemes change rules help: A configurable approach to coreference resolution beyond OntoNotes. In *Proceedings of the NAACL2016 Workshop on Coreference Resolution Beyond OntoNotes (CORBON)*, pages 92–101.


## A Transcription