Is one head enough?
Mention heads in coreference annotations compared with UD-style heads

Anna Nedoluzhko
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
nedoluzhko@ufal.mff.cuni.cz

Michal Novák
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
mnovak@ufal.mff.cuni.cz

Martin Popel
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
popel@ufal.mff.cuni.cz

Zdeněk Žabokrtský
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
zabokrtsky@ufal.mff.cuni.cz

Daniel Zeman
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
zeman@ufal.mff.cuni.cz

Abstract

We present an empirical study that compares mention heads as annotated manually in four coreference datasets (for Dutch, English, Polish, and Russian) on one hand, with heads induced from dependency trees parsed automatically, on the other hand. For parsing, we used UDPipe 2.6, a modern parser trained using the Universal Dependencies collection. We show that majority of mismatches (64%–94%) can be attributed to several classes of systematic differences in how the notion of head is treated in the respective data resources, while mismatches caused by parsing errors are relatively rare (4%–15%). Our conclusion is that consistency would be gained in (and across) coreference resources after migration to UD-style mention heads, without losing substantial information. This can be achieved with sufficient accuracy using modern dependency parsers even for coreference corpora that lack manual head annotation.

1 Introduction

Coreference is a relation between expressions in a text which refer to the same real-word entity or event; the referring expressions are called *mentions*. In most datasets annotated with coreference relations (see Nedoluzhko et al. (2021) for a survey), a mention is represented simply by specifying the corresponding sequence of tokens (called a mention *span*), typically contiguous, mostly belonging to a single sentence.

Naturally, a mention span can be analyzed syntactically. There is a vague consensus that some tokens, often delimited syntactically, carry more important information from the coreference resolution perspective than other tokens. The most crucial part is called a *minimum span* by some (as opposed to *maximum span* denoting the whole span; see e.g. Uryupina et al. (2020), Hirschman and Chinchor (1998)), or simply a *head* by others (Ogrodniczuk et al., 2013), which is the term we adhere to. Identifying a mention's head is motivated not only linguistically, but also technically: with a long-span mention, there is a higher risk of annotation noise and requiring the exact match when evaluating span boundary prediction could be misleading. See e.g. Uryupina et al. (2020), Elsner and Charniak (2010), Peng et al. (2015), or Wiseman et al. (2016) for more arguments on the importance of head for the task of coreference resolution.

The notion of mention head in coreference annotations largely resembles the notion of head in dependency treebanks; however, with a few exceptions such as the Prague Dependency Treebank (Hajič et al.,

2020) (PDT for short), the coreference and dependency-treebanking annotation efforts remain isolated to a surprising degree.

In this paper we present a novel empirical study that compares manually annotated mention heads within coreference annotation projects with syntactic heads identified automatically by a modern parser trained on dependency treebanks from the Universal Dependencies (UD) collection (de Marneffe et al., 2021). Our long-term motivation is based on the expectation that making the mention head notion convergent with heads induced from dependency structures following the UD guidelines could result in (a) improved annotation consistency in existing coreference datasets, and (b) more efficient and faster development of new coreference datasets (e.g. because of possible reuse of UD-related software tools), especially when it comes to extensions to multiple languages. However, in a shorter-term perspective, we should first try to explain the nature of differences between mention heads as annotated in existing coreference datasets on the one hand and UD-compliant heads of mentions on the other.

We make use of the CorefUD 0.2 collection, which contains 17 coreference datasets for 11 languages converted to a common annotation scheme (Nedoluzhko et al., 2021). There is some notion of head used explicitly or implicitly in 13 out of the 17 datasets. However, we limit ourselves only to datasets in which mention heads are marked explicitly, and, at the same time, whose coreference annotations were created without using full-fledged hand-annotated syntactic structures (dependency or constituency). Thus, for example, the Prague Dependency Treebank dataset is excluded, since coreference and dependency annotations are tightly connected in it by design. The selection criterion leads to four resources: ARRAU (Uryupina et al., 2020) for English, COREA for Dutch (Hendrickx et al., 2008), Polish Coreference Corpus (Ogrodniczuk et al., 2015; Ogrodniczuk et al., 2013), and Russian Coreference Corpus (Toldova et al., 2014). Datasets in CorefUD 0.2 have been parsed using the UDPipe 2 tool (Straka, 2018) with very recent parser models.

The rest of the paper is structured as follows. Section 2 summarizes different approaches to the notion of head, both from the syntactic and coreference perspectives. Section 3 gives basic information about the four coreference data resources included in our study. Section 4 describes our annotation of mentions selected from the four resources; we focused on mentions in which the mention head marked in the original coreference resource does not match the root of the mention in terms of automatically parsed UD tree. Section 5 analyzes and exemplifies types of such mismatches. Finally, Section 6 concludes.

2 Related work

2.1 Head in dependency annotation schemes

One can easily foresee that – in spite of the recent progress in parsing technology – there will be non-negligible amount of head mismatches which are due to parsing errors; similarly, a non-zero amount of errors in manual annotation of mention heads can be expected too. However, we are interested in more principled sources of variability of the notion of head.

It was recognized by dependency-oriented scholars long time ago that multiple types of dependencies may be distinguished (especially syntactic and semantic ones), and that syntactic dependencies should not be confused with other types of relations, see e.g. a discussion on "double dependency" and "mutual dependency" in Mel'čuk and others (1988). However, the trend in the current dependency treebanking in the last decade is inclined rather to maximize simplicity and robustness, with Universal Dependencies (de Marneffe et al., 2021) being the most prominent representative, rather than to design multilayered annotation schemes with strictly separated hypotactic and paratactic "brackets" (with the latter ones possibly interpreted as additional "dimensions" of dependency trees (Sgall, 1998)) on each layer. This trend has a clear rationale especially if quick portability to multiple languages is one of the modern priorities, however, on the other hand, such formally simple structures are prone to various confusions concerning the notion of head.

The fact that in some cases there is no unique obvious way for choosing a head of an expression, has been noticed many times as it has inevitable practical consequences in dependency-oriented projects. Above all, annotators' intuition concerning the dependency structure of sentences is insufficient for reaching reasonable annotation consistency, and thus artificial annotation rules must be introduced by

convention. This can be illustrated by extensive annotation guidelines developed basically in every mature dependency treebanking project. We believe that most of the observed variability in the notion of head can be attributed to the following sources, as discussed in more detail the subsections below:

- opposite direction of syntactic and semantic dependencies (and other non-parallelisms),
- representing functional words as nodes of their own,
- representing paratactic relations within dependency trees,
- no obvious head-dependent asymmetry in a syntactic constituent.

2.1.1 Opposite direction of syntactic and semantic dependencies

Several types of constructions are recognized in literature in which the direction of a syntactic dependency relation manifested by overt surface morphosyntactic means (such as agreement) is opposite to what is considered as semantic dependency; the syntactic and semantic heads are swapped, in other words.

When designing treebank annotation guidelines, the authors either have to indicate whether syntactic or semantic dependencies are the preferred ones, or, alternatively, provide technical means for capturing both. The latter option can be illustrated by PDT, in which there are two separate dependency trees, one of them capturing surface syntax and the other one capturing deep syntax and semantics (to some extent). Similarly, the Enhanced representation in Universal Dependencies (Nivre et al., 2020, Section 3.4) adds extra edges to make explicit some semantically relevant relations that are otherwise implicit in basic dependencies.

2.1.2 Functional words

If functional words have nodes of their own in a dependency representation, it can lead to problems related to head choice. A functional word is usually clearly associated with an autosemantic (meaningful) word, however, it is not clear which of them should be the head (more precisely, either choice can be justified with reasonable arguments, and one simply has to choose). Examples of such pairs are a preposition and a noun in a prepositional group, an auxiliary verb and an autosemantic verb in a complex verb form, or a determiner and a noun. For instance, if a prepositional group is considered, PDT surface syntax guidelines make the preposition the governor and put the noun below, while the two are connected the other way round in UD. If an auxiliary verb in a complex verb form bears congruent categories, then it becomes the governor in the PDT, while the autosemantic component of the complex verb form is the governor in most cases. Both PDT and UD annotation styles attach determiners below nouns being determined, but determiners are treated as governors of noun phrases in the Danish Dependency Treebank (Kromann et al., 2003).

A more complex example is that of expletives: in some cases insertion of expletive expressions (such as pronouns) is needed or preferred in a language, for instance if valency of a matrix clause verb requires a morphological case to be manifested with its argument, but the argument is a subordinating clause. Then, again, it is not clear whether the expletive pronoun or the subordinating clause head should be chosen as the head of it all.

2.1.3 Paratactic structures

In the case of parataxis, two syntactically connected expressions are in an equal relation with each other, instead of being subordinated one to the other. In other words, there is no head-dependent asymmetry. Typical examples are coordination and apposition constructions. Especially coordination has always been a nightmare for dependency grammarians, as it is very frequent and interferes in various ways with dependency relations. However, as long as we preserve the design decision that all we have for syntactic representation is nodes and edges, we have to encode paratactic constructions in this way too. There is a surprising number of different possible encodings for doing so, and a smaller, but even more surprising number of encodings that has been really used in existing treebanks, see Popel et al. (2013) for a survey. However, in most cases it boils down to either using coordination conjunction as the head node, or using one of the conjuncts as the head, selected in some canonical way.

2.1.4 No overt head-dependent asymmetry

Besides paratactic structures, there are also other types of expressions in which we perceive some internal structure and for which we do not possess intuition about what should be the head, but which are not paratactic either. A frequent example is a personal name consisting of a given name and a family name. UD has a dedicated relation type, `flat`, which is used in such exocentric constituents; the first word serves as the technical head, but there is no claim that it is a syntactically (or semantically) motivated head.

To summarize, head choice is far from obvious in various cases, which has both deeply linguistic and purely technical reasons; such situations can only be resolved unambiguously by adhering to artificial annotation rules.

2.2 Head in coreference annotation schemes

For a better orientation, we suggest to classify language data resources containing coreference annotation tentatively as follows:

- head-agnostic approaches,
- head-aware approaches,
- head-centric approaches.

2.2.1 Head-agnostic approaches.

In head-agnostic approaches, a mention is considered the only meaningful unit that is needed for annotating coreference relations and no attempt to find its internal structure is made (at least not to our knowledge).

Examples of head-agnostic approaches are Potsdam Commentary Corpus (Bourgonje and Stede, 2020), the English-German parallel coreference corpus ParCorFull (Lapshinova-Koltunski et al., 2018), and Lithuanian Coreference Corpus (Žitkus and Butkienė, 2018).

2.2.2 Head-aware approaches.

In head-aware approaches, a mention delimited as a sequence of tokens is still the main entity, however, its internal structure is analyzed syntactically[1] (completely or partially) and/or its head is marked explicitly.

Examples of head-aware approaches are Spanish and Catalan data contained in AnCora (Recasens and Martí, 2010) and English data contained in ARRAU (Uryupina et al., 2020).

2.2.3 Head-centric approaches.

In head-centric approaches, it is the head of a mention that is considered to be the argument of a coreference relation, while the exact span in terms of a token sequence is less important (or even left underspecified). Coreference datasets from the PDT family, in which coreference relations connect tectogrammatical (deep-syntactic) nodes and mention span is defined only implicitly, are examples of this approach.

Examples of head-centric approaches are the Prague Dependency Treebank (Hajič et al., 2020) and the Prague Czech-English Dependency Treebank (Nedoluzhko et al., 2016).

3 Coreference datasets with hand-annotated mention heads

Our analysis is based on four datasets from CorefUD 0.2[2] whose original source corpora contain manual annotation of mentions: ARRAU, Polish Coreference Corpus, COREA, and Russian Coreference Corpus.

3.1 ARRAU

The ARRAU Corpus of Anaphoric Information (Uryupina et al., 2020) (further abbreviated as English-ARRAU) is a multi-genre corpus of English which provides large-scale annotations of a wide range of anaphoric phenomena. In English-ARRAU, the special attribute MIN (or minimal span) is manually annotated, similarly as it was once decided for MUC-7 (Hirschman and Chinchor, 1998). This attribute

[1] An analysis whether or not coreference mentions do correspond to subtrees of UD trees can be found in Popel et al. (2021), without a special attention paid to heads, though.

[2] `http://hdl.handle.net/11234/1-4598`

corresponds to the head noun for non-proper nominal mentions, or to the entire proper name (for example, first name and surname) in case of multi-word named entities. It is not explicitly stated in the guidelines, if syntactic or semantic heads are preferred. According to the MUC-7 coreference task definition[3], it maybe deduced that syntactic heads are preferred. However, this has not been stated explicitly for MUC-7 neither.

3.2 Polish Coreference Corpus

The Polish Coreference Corpus (Ogrodniczuk et al., 2013; Ogrodniczuk et al., 2015) (further abbreviated as Polish-PCC) is a corpus of Polish nominal coreference built upon the National Corpus of Polish. In Polish-PCC, semantic heads, i.e. the most important words from the point of view of the mention's sense, are annotated. The semantic head of a typical nominal group corresponds to the syntactic head but there are some exceptions. For example, in numeral groups like *duzo pieniedzy* 'a lot of money', or *trzech z was* 'three of you', the numeral is the syntactic head, and the noun is the semantic head and is annotated as such in Polish-PCC. The reason for such decision is the claim that coreference is a phenomenon on the level of semantics and discourse more than on the syntactic level. Thus, understanding the semantically central elements should help establish discourse links. Although not explicitly found in the guidelines, the head is understood semantically (an item with larger semantic weight is annotated as head) also in other types of constructions (*od 1999 roku* 'from the year 1999' with the numeral as a head, *pan Ziolkowski* 'Mr. Ziolkowski' with the surname as a head, etc.).

3.3 COREA

The COREA coreference corpus (Hendrickx et al., 2008) (further abbreviated as Dutch-COREA) is a collection of written and transcribed oral texts in Dutch annotated for creating a coreference resolution system. Mentions are strings of text with specially distinguished heads which are defined as minimum strings representing semantic heads of the constituents. Nevertheless, rather than annotated from scratch, the semantic heads were acquired by manual post-editing of the heads obtained from syntactic representation of the underlying texts. For multi-word named entities, the head includes all words of the corresponding entity.

3.4 Russian Coreference Corpus

Russian Coreference Corpus (Toldova et al., 2014) (further abbreviated as Russian-RuCor) is annotated with anaphoric and coreferential relations between noun groups. Mentions are annotated as linear spans, with additionally distinguished heads. Similarly as for English-ARRAU and Dutch-COREA, heads are defined as one-word syntactic heads for common nouns and as sequences of words for multi-word proper nouns. For 'common noun + proper noun' constructions like *the Pushkin street*, the guidelines require the whole multi-word sequences to be annotated as heads, but in the annotated data, only one word is chosen as head (mostly the proper noun).

The comparison of the guidelines for head annotation in the resources under analysis shows that there are differences in the following aspects:

- *Syntactic or semantic understanding of heads*: Semantic heads are explicitly claimed to be annotated in Polish-PCC and partly in Dutch-COREA; in English-ARRAU and Russian-RuCor, there is no explicit claim about the syntactic nature of annotated heads but it may be deduced from the guidelines examples;
- *Possibility to annotate multi-word entities as a head*: Possible for multi-word named entities in English-ARRAU, Russian-RuCor and Dutch-COREA and not applied in Polish-PCC;
- *Choice of the head in 'common noun + proper noun' constructions*: the proper name in English-ARRAU, Dutch-COREA and Polish-PCC and both entities in Russian-RuCor;

[3] `https://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html`

	count	[%]				
				annotated head		
CorefUD dataset	all	one-word	non-catena	missing	same	different
Dutch-COREA	26,476	38.9	2.7	4.6	47.2	6.6
English-ARRAU	57,681	30.0	5.4	3.1	56.3	5.3
Polish-PCC	150,706	49.1	5.0	0.1	44.3	1.5
Russian-RuCor	12,632	68.9	1.1	0.1	27.3	2.5

Table 1: Statistics on mentions in the whole dataset. *all* is the total number of mentions in the train section of a given dataset. The other columns show percentage breakdown into mention types described in the first paragraph of Section 4. The types are detected automatically in a given order, so e.g. a non-catena mention with no annotated head is assigned the non-catena type (not missing head). The last column shows a percentage of multi-word catena mentions with a mismatch in annotated and syntactic head; a sample of 100 mentions of this type was annotated as shown in Table 2.

- Different technical conventions for apposition and coordination structures, special construction with dollar, percent, etc.

Dependency trees in the four datasets under discussion have been obtained for CorefUD 0.2 using UDPipe 2 and models trained on UD 2.6, namely on English-GUM (Zeldes, 2017), Polish-LFG (Patejuk and Przepiórkowski, 2018), Dutch-LassySmall (Bouma and van Noord, 2017), and Russian-SynTagRus (Droganova et al., 2018).

4 Annotation of head mismatches

In our study, we focus on mentions, in which the head coming from the original annotation differs from the head of the mention with respect to the tree produced by automatic dependency parsing.[4] Consequently, all one-word mentions are excluded. In addition, we take into consideration only such mentions whose inner dependency structure forms a *catena*, i.e. a connected subgraph of a dependency tree (Osborne et al., 2012).[5] Moreover, we focus only on mentions where at least one head was annotated.[6]

We randomly sampled 100 such mentions from a train section of each of the four CorefUD datasets under analysis. The examples were examined and annotated by the authors of this work. As none of us is a speaker of Dutch, we utilized public machine-translation services in order to understand the example sentences.

During the annotation process, we settled upon the following categories of head mismatches:

- WRONG – we consider the mismatch to be an error.
  - WRONGTREE – the automatically parsed UD tree is wrong
  - WRONGSPAN – wrong syntactic head caused by a wrong mention span, usually due to extra tokens.
  - WRONGHEAD – the manual annotation of head is wrong, i.e. it does not follow the original project annotation guidelines (or at least we were not able to find any guideline which would support such head annotation).

[4]Multiple words could be annotated as heads (or minimal span) in the original annotation. In such cases, we focus on mentions where the syntactic head is not among the set of annotated heads.

[5]Note that catena differs from a **subtree**, which is a catena that spans the head and **all** its descendants. Non-catena mentions have multiple nodes that can be considered syntactic heads of the mention (i.e. their dependency parent is not part of the mention).

[6]See Table 1 for statistics on the total count of mentions and their breakdown into the abovementioned types excluded from the annotation (one-word, non-catena, missing-head, same-head).

	OK				WRONG		
CorefUD dataset	COORD	FLAT	NUM	OTHER	TREE	SPAN	HEAD
Dutch-COREA	25	31	11	7	7	7	16
English-ARRAU	1	44	14	13	4	0	25
Polish-PCC	11	21	23	9	15	1	13
Russian-RuCor	0	85	7	2	5	0	1

Table 2: Result of our annotation of differences in annotated and syntactic heads in a sample of 100 mentions in each dataset. Disclaimer: Individual cases of WRONGHEAD may turn out to be cases of OK or vice versa. A deeper analysis of such cases is a subject of future studies.

- OK – the mismatch in head annotations is correct, as the respective guidelines do not agree on the head for a given phenomenon
  - OK-COORD – the first conjunct of a coordination is always marked as a head in UD. The original annotation marks the coordination conjunction or another conjunct as a head, instead. This is an example of the parataxis (see Section 2.1.3).
  - OK-FLAT – UD chooses the first token as head in flat structures (such as names, marked with deprel `flat`), appositions (marked with deprel `appos`) and lists, while the original dataset annotators decided to analyze it as a non-flat structure. This is an example of non-overt head-dependent asymmetry as we describe it in Section 2.1.4, and parataxis for apposition (see Section 2.1.3).[7]
  - OK-NUM – the mismatch is caused by an opposite direction of syntactic and semantic dependencies (see Section 2.1.1). This most often includes numerals and containers (e.g. *a group of people*).
  - OK-OTHER – another subtype of OK.

5 Analysis and discussion

Table 2 summarizes the head mismatches annotation in the selected datasets. As we can see, there is a relatively low number of mismatches caused by wrong parsing. With a slightly larger number of such cases in Polish-PCC, there are just up to 7% of wrongly parsed annotated mentions in English-ARRAU, Dutch-COREA and Russian-RuCor. One of the reasons is that we included only multi-word catena structures into the analysis.[8] The remaining cases of wrong parsing are specific syntactic or derivation constructions, e.g. the deadjectival noun *teściowa /mother-in-law/* in Example 1[9] from Polish which is falsely recognized as an adjective in UDPipe and thus gets a dependent position in the parsed tree. The surprisingly low overall number of parse errors can be justified by comparative simplicity of parsing of noun phrases (the majority of mentions are noun phrases).

[7] Even though appositions are paractic constructions, we rather included them in the OK-FLAT category. The reason is that they are closely related to hypotactic constructions such as *president Trump*, which are in fact treated as appositions in some of the datasets (e.g. Dutch-COREA).

[8] Note that Polish-PCC has the lowest percentage of head mismatches according to the last column in Table 1. Thus, in Polish-PCC we could expect $1.5\% \cdot 15\% = 0.23\%$ mentions with head mismatch caused by wrong parsing in the whole dataset, while in Dutch-COREA it is twice as much: $6.6\% \cdot 7\% = 0.46\%$.

[9] Examples in this work are presented in both glosses and trees. The first line of the gloss shows the original sentence / excerpt / phrase, optionally followed by its word-to-word translation and smooth translation to English. Nodes in the dependency tree show the word form, part-of-speech tag and dependency relation to the node's parent. While in gloss the annotated mention is typeset in bold, `**token**` is used to mark each token of the mention in the tree. The annotated mention head and syntactic head given by the parser are labelled only in the tree.

(1) *Przypomniał* *sobie* *pazury* ***teściowej*** ***doktora*** ***Goldmanna***.
He remembered himself claws of mother-in-law of dr. Goldmann.
'He remembered Dr. Goldmann's mother-in-law's claws.'

```
Przypomniał VERB root
  sobie PRON iobj
  pazury NOUN obj
    **teściowej ADJ amod <======= annot. head**
    **doktora NOUN nmod:poss <----- synt. head**
    **Goldmanna PROPN flat**
  . PUNCT punct
```

Another apparent general observation is a disproportion of incorrectly parsed or annotated sentences (WRONG labels) between Russian-RuCor (6%) and the other datasets (28–30%). This is likely a consequence of annotation mismatches in proper nouns that prevail in the selected sample (see Section 5.2).

Our analysis reveals a number of mismatches (i) between syntactic heads generated by the UD parser and manually annotated heads in the datasets, but also inconsistencies (ii) across the datasets and (iii) within the annotated datasets themselves. The most typical categories of mismatches (OK-COORD, OK-FLAT and OK-NUM) and annotation inconsistencies are addressed in the following subsections.

5.1 Heads in coordinations

The prevailing reason for mismatches of the OK-COORD type is that the coordination conjunction is annotated as the head of a coordination mention. In total it accounts for 73%.

In cases where a non-first conjunct is annotated as a mention head, the conjunct often comprises information that is shared among all conjuncts, e.g. in Example 2 from Dutch-COREA.

(2) ***gezonde*** ***bacterie-*** ***of*** ***virusdragers***
healthy bacteria or virus carriers
'healthy bacteria or virus carriers'

```
  **gezonde VERB amod**
**bacterie- X obl <------------- synt. head**
  **of CCONJ cc**
  **virusdragers NOUN conj <==== annot. head**
```

5.2 Heads in expressions with proper names

Constructions with proper names form a great deal of the OK-FLAT category. There are different annotation conventions for annotating heads in such constructions in UD (de Marneffe et al., 2021) and across the annotated datasets (see descriptions of the datasets in Section 3). Generally, whereas in phrases like *President Carter* the annotator more often chooses the proper noun as head (because it is referentially concrete), it is the first word (i.e. the general name *President* in our example) according to the UD convention (see Example 3 from English-ARRAU).

(3) *Mr. Hastings was appointed to the federal bench by* ***President Carter***

```
  Mr. PROPN nsubj:pass
    Hastings PROPN flat
  was AUX aux:pass
appointed VERB root
  to ADP case
  the DET det
  federal ADJ amod
  bench NOUN obl
    by ADP case
  **President PROPN obl <------ synt. head**
    **Carter PROPN flat <====== annot. head**
```

Interestingly, there are 85% such cases in the Russian sample, although the guidelines rather advise to label both expressions as a multi-word head. See the expression *журнала Time /Time magazine/* in Example 4, where the annotated mention head is the proper name and the UD head is the common noun *журнала /magazine/*. This type of mismatches is also frequent in other datasets, see e.g. *moja babcia Zofia /my grandma Zofia/* in Example 5 (Polish-PCC) or *vitamine C* in Example 6 (Dutch-COREA).

There is also a non-negligible number of inconsistencies in annotation of multi-word named entities within the datasets. Although, the guidelines require to mark the entire multi-word units as heads in all datasets except Polish-PCC, in some cases, only one more semantically significant word is annotated. See the annotation of only surname in the multi-word name Hillary Clinton in Example 7 from Russian-RuCor.

(4) *Подписка на «планшетную» версию* ***журнала*** ***Time***
Subscription to "tablet" version of magazine Time
'Subscription to the "tablet" version of Time magazine'

```
Подписка NOUN nsubj
  на ADP case
    « PUNCT punct
  планшетную ADJ amod
    » PUNCT punct
версию NOUN nmod
  **журнала NOUN nmod <---------- synt. head**
    **Time PROPN flat:foreign <=== annot. head**
```

(5) ***moja babcia Zofia Gołąbowa*** *mieszkała w kamienicy*
my grandma Zofia Gołąbowa lived in tenement house
'my grandmother, Zofia Gołąbowa, lived in a tenement house'

```
  **moja DET det**
  **babcia NOUN nsubj <------ synt. head**
    **Zofia PROPN flat <===== annot. head**
    **Gołąbowa PROPN flat**
mieszkała VERB root
  w ADP case
  kamienicy NOUN obl
```

(6) *door* ***vitamine C*** *of amandelzuur*
by vitamin C or mandelic acid
'by vitamin C or mandelic acid'

```
  door ADP case
**vitamine NOUN nmod <---- synt. head**
  **C PROPN appos <======= annot. head**
    of CCONJ cc
  amandelzuur NOUN conj
```

(7) *Запрос был подан* ***госсекретарю*** ***США*** ***Хиллари*** ***Клинтон***
request was submitted secretary USA Hillary Clinton
'The request was submitted to the US Secretary Hillary Clinton'

```
  запрос NOUN nsubj:pass
  был AUX aux:pass
подан VERB root
  **госсекретарю NOUN iobj <--------- synt. head**
    **США PROPN nmod**
    **Хиллари PROPN appos**
      **Клинтон PROPN flat:name <==== annot. head**
  . PUNCT punct
```

5.3 Heads in expressions with numerals and quantifiers

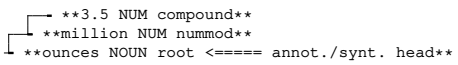
Head mismatches in constructions with numerals grouped under the OK-NUM category may be further divided into the following subgroups.

Cardinal numerals. Numeral mentions where a noun-like word is modified by a number (e.g. *five cars*) are typical cases of head mismatches. In most of them, the modified word is in fact a currency's name or symbol (e.g. *$25 million*, *vijfhonderd zestig miljoen gulden /five hundred and sixty million guilders/*, *90 млрд рублей /90 billion rubles/*).

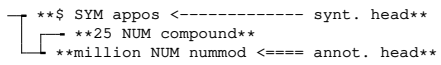
Heads are annotated inconsistently across datasets: numerals prevailingly serve as heads in English-ARRAU (Examples 8–9) and Polish-PCC (Example 10), while it is the modified words in Russian-RuCor (Example 13). Nevertheless, annotation of mention heads seems to be inconsistent also within some of the datasets. Let us look into mentions with a *$* symbol as their syntactic head (*$* mention) in English-ARRAU. Out of 727 such mentions scattered over 179 original documents, only in 43% of them the annotated head (minimal span) matches the syntactic head. Interestingly though, if an original document contains a matching *$* mention, on average more than 92% of all *$* mentions in the document are matching, too. The observed inconsistency thus occurs rather across than within original documents, suggesting that it is an artifact of the annotation workload having been distributed among multiple annotators on the document level.

The mismatches between syntactic and annotated heads partly result also from inconsistencies in parses. However, we do not categorize them as parsing errors (WRONGTREE) since UDPipe models almost perfectly mimic the inconsistency that can be seen already across the manually annotated UD subcorpora they were trained on. While syntactic annotation of single-token numerals (e.g. *five cars*) seem to be identical across the languages, it differs considerably for multi-token numerals with large number names such as thousands, millions etc. (cf. Examples 8-13). Moreover, in Russian-SynTagRus the tree of multi-token numerals is shaped differently based on whether the word representing the large number name is in singular (Example 12) or plural (Example 13).

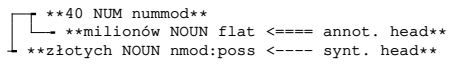
(8) ***3.5 million ounces***



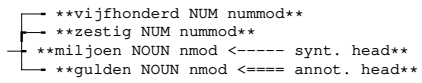
(9) ***$25 million***



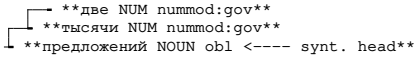
(10) ***40 milionów złotych***
40 million złoty



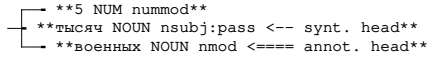
(11) ***vijfhonderd zestig miljoen gulden***
five hundred sixty million guilders



(12) ***две тысячи предложений***
two thousand.GEN.SG sentences

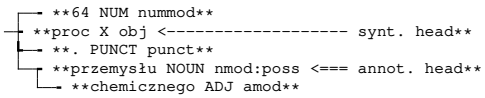


(13) ***5 тысяч военных***
5 thousand.GEN.PL soldiers

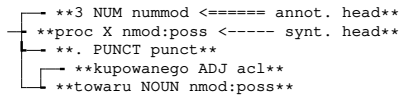


Syntactically governing numerals and containers. In constructions with governing numerals (e.g. *one of the candidates*, *all of this*) and so-called 'containers' (e.g. *group of tourists*), UDPipe systematically marks the numerals or containers as heads. On the other hand, manual annotation often chooses their syntactic dependent members as more important, putting the emphasis on the semantic point of view. Nevertheless, Examples 14–15 from Polish-PCC and Examples 16–17 from Russian-RuCor show that the manual annotation of mention heads in constructions with containers and governing numerals, respectively, is not systematic. Although we admit there may be another aspect (e.g. semantic salience) that convinced the annotators to label heads in these examples differently, it is neither obvious nor described in the guidelines.

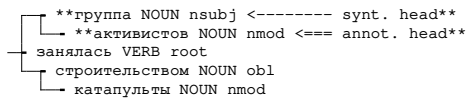
(14) ***64 proc. przemysłu chemicznego***
64 perc. industry chemical
'64% chemical industry'



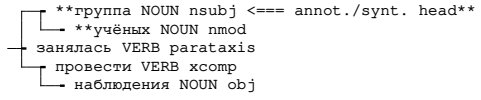
(15) ***3 proc. kupowanego towaru***
3 perc. purchased goods
'3% purchased goods'



(16) ***группа активистов*** *занялась строительством*
group activists.GEN took up construction
катапульты
catapult.GEN
'a group of activists took up the construction of the catapult'

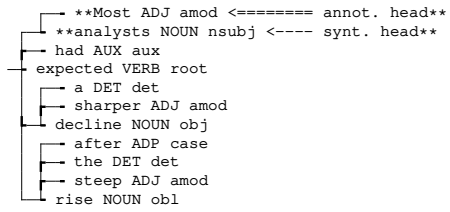


(17) ***группа учёных*** *планировала провести наблюдения*
group scientists.GEN planned to conduct observations
'a group of scientists planned to conduct observations'

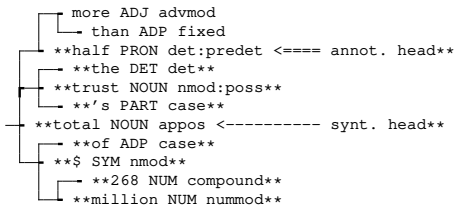


Quantifiers as determiners. Interestingly, we find quite a lot of cases of quantifiers in the syntactic position of determiners (*some, most, each, half* and even *no*). They are heads neither from the syntactic nor the semantic point of view. However, in some cases they are marked as heads in manual annotations, e.g. in *most analysts* in Example 18, *half the total* in Example 19, *some investors*, *each bond*, *no trading* (all from English-ARRAU) and in *несколько серых пятен /some grey spots/* from Russian-RuCor.

(18) ***Most analysts*** *had expected a sharper decline after the steep rise*



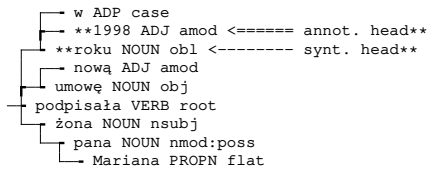
(19) *more than* ***half the trust's total of $268 million***



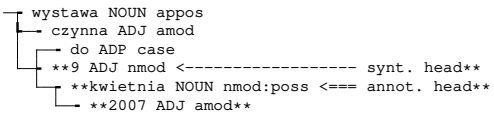
The reasons for such mismatches may be twofold. First, these constructions are not clearly distinct from the structures like 'most of people, 'half of people' where *most* and *half* are syntactic heads. Another reason may be higher salience of the determiners in the given contexts.

Dates. Mismatches in dates seem to appear only in Polish-PCC. Years and months (if present) are consistently annotated as mention heads, as illustrated in Examples 20 and 21, respectively. Therefore, it should not be too difficult to obtain such mention heads using a rule-based transformation based on syntax.

(20) *w* ***1998*** ***roku*** *nową umowę podpisała żona pana Mariana*
in 1998 year new contract sign wife Mr. Marian's
'in 1998, a new contract was signed by Mr. Marian's wife '



(21) *wystawa czynna do* ***9*** ***kwietnia*** ***2007***
exhibition open until 9 April 2007
'the exhibition is open until April 9, 2007'



6 Conclusion

We have provided a novel comparison of syntactic dependency structure on one hand, and annotation of coreferential mentions on the other hand. In particular, we focus on the notion of mention heads in coreference datasets where such a notion exists and it is not designed to be identical to the syntactic head. Nevertheless, we can compare mention heads with syntactic heads thanks to the CorefUD collection, which contains coreference corpora with dependency structures predicted by the UDPipe parser. We collected mention instances where the syntactic head did not match the designated mention head, then we manually examined a subset of such instances and analyzed the likely reasons for the difference.

If we summarize our observations, the UD heads and manually annotated mention heads coincide in majority of multi-token mentions in all four studied datasets already now, while most differences can be attributed to one of the following reasons:

- heads of a mention are different because of an error made by the UD parser, or because of an error made by an annotator; the amount of parsing errors is surprisingly low, likely due to relative simplicity of parsing of noun phrases (and will hopefully further fade out with progress in parsing technology),
- heads are selected using rather technical than linguistic rules in expressions such as named entities or coordination structures (in which linguistic intuitions for heads are weak); rule-based transformations could be used for translating UD convention to a coreference dataset convention or *vice versa*,
- semantic rather than syntactic heads are chosen in coreference annotations, e.g. in expressions with numerals; however, with an exception of some types of expressions (e.g. 'containers'), again a few rule-based patterns on the UD tree of a mention could be used to automatically identify the semantic head,

- in some cases, mention head annotations in coreference datasets bear information that seems intuitively semantically salient (such as contrast) and undeducible from UD syntax; however, such cases are rare and typically not supported by coreference annotation guidelines.

Let us conclude by answering the question from the title. It seems that both inter-project and intra-project consistency would be gained and almost nothing would be lost if we start adhering to the UD notion of heads in mentions in coreference projects, instead of annotating coreference-specific heads. In addition, quality of mention heads derived from automatic UD parses based on modern parsing technology is quite high, which would further reduce potential benefits of manual annotation of mention heads in future coreference-oriented projects.

Acknowledgements

This work was supported by the Grants GA19-14534S and 20-16819X (LUSyD) of the Czech Science Foundation; LM2018101 (LINDAT/CLARIAH-CZ) of the Ministry of Education, Youth, and Sports of the Czech Republic; and EC/H2020/825303 (Bergamot) of the European Commision.

We thank the three anonymous reviewers for their very insightful and useful comments.

References

Gosse Bouma and Gertjan van Noord. 2017. Increasing return on annotation investment: The automatic construction of a Universal Dependency treebank for Dutch. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 19–26, Gothenburg, Sweden, May. Association for Computational Linguistics.

Peter Bourgonje and Manfred Stede. 2020. The Potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1061–1066, Marseille, France, May. European Language Resources Association.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 53–66, Linköping, Sweden. Linköping University Electronic Press.

Micha Elsner and Eugene Charniak. 2010. The same-head heuristic for coreference. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 33–37.

Jan Hajič, Eduard Bejček, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague Dependency Treebank - Consolidated 1.0. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 5208–5218, Marseille, France. European Language Resources Association.

Iris Hendrickx, Gosse Bouma, Frederik Coppens, Walter Daelemans, Veronique Hoste, Geert Kloosterman, Anne-Marie Mineur, Joeri Van Der Vloet, and Jean-Luc Verschelde. 2008. A coreference corpus and resolution system for Dutch. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Lynette Hirschman and Nancy Chinchor. 1998. Appendix F: MUC-7 coreference task definition (version 3.0). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

Matthias T. Kromann, Line Mikkelsen, and Stine Kern Lynge. 2003. Danish Dependency Treebank. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, pages 217–220.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: A parallel corpus annotated with full coreference. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Igor Aleksandrovič Mel'čuk et al. 1988. *Dependency syntax: theory and practice*. SUNY press.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. 2016. Coreference in Prague Czech-English Dependency Treebank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 169–176, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, and Daniel Zeman. 2021. Coreference meets Universal Dependencies – a pilot experiment on harmonizing coreference datasets for 11 languages. Technical Report 66, ÚFAL MFF UK, Praha, Czechia.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 4034–4043, Marseille, France. European Language Resources Association.

Maciej Ogrodniczuk, Katarzyna Glowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2013. Polish coreference corpus. In *Human Language Technology. Challenges for Computer Science and Linguistics - 6th Language and Technology Conference, LTC 2013, Poznań, Poland, December 7-9, 2013. Revised Selected Papers*, volume 9561 of *Lecture Notes in Computer Science*, pages 215–226. Springer.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2015. *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter.

Timothy Osborne, Michael Putnam, and Thomas Groß. 2012. Catenae: Introducing a novel unit of syntactic analysis. *Syntax*, 15(4):354–396.

Agnieszka Patejuk and Adam Przepiórkowski. 2018. *From Lexical Functional Grammar to Enhanced Universal Dependencies: Linguistically informed treebanks of Polish*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Haoruo Peng, Kai-Wei Chang, and Dan Roth. 2015. A joint framework for coreference resolution and mention head detection. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 12–21.

Martin Popel, David Mareček, Jan Štěpánek, Daniel Zeman, and Zdeněk Žabokrtský. 2013. Coordination structures in dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 517–527, Sofija, Bulgaria. Association for Computational Linguistics.

Martin Popel, Zdeněk Žabokrtský, Anna Nedoluzhko, Michal Novák, and Daniel Zeman. 2021. Do UD Trees Match Mention Spans in Coreference Annotations? In *Findings of EMNLP 2021*. Association for Computational Linguistics.

Marta Recasens and M. Antònia Martí. 2010. AnCora-CO: Coreferentially Annotated Corpora for Spanish and Catalan. *Lang. Resour. Eval.*, 44(4):315–345, December.

Petr Sgall. 1998. Teorie valence a její formální zpracování. *Slovo a slovesnost*, 59(1):15–29.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, October. Association for Computational Linguistics.

Svetlana Toldova, Anna Roytberg, Alina Ladygina, Maria Vasilyeva, Ilya Azerkovich, Matvei Kurzukov, G. Sim, D.V. Gorshkov, A. Ivanova, Anna Nedoluzhko, and Yulia Grishina. 2014. Evaluating Anaphora and Coreference Resolution for Russian. In *Komp'juternaja lingvistika i intellektual'nye tehnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii Dialog*, pages 681–695.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU Corpus. *Natural Language Engineering*, 26(1):95–128, January.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2016. Learning Global Features for Coreference Resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004, San Diego, California, June. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3):581–612, September.

Voldemaras Žitkus and Rita Butkienė. 2018. Coreference annotation scheme and corpus for Lithuanian language. In *Fifth International Conference on Social Networks Analysis, Management and Security, SNAMS 2018, Valencia, Spain, October 15-18, 2018*, pages 243–250. IEEE.