CaMEL: Case Marker Extraction without Labels 🐪

Leonie Weissweiler[*], **Valentin Hofmann**[†*], **Masoud Jalili Sabet**[*], **Hinrich Schütze**[*]

[*]Center for Information and Language Processing, LMU Munich
[†]Faculty of Linguistics, University of Oxford
{weissweiler,masoud}@cis.lmu.de
valentin.hofmann@ling-phil.ox.ac.uk

Abstract

We introduce **CaMEL** (**Ca**se **M**arker **E**xtraction without **L**abels), a novel and challenging task in computational morphology that is especially relevant for low-resource languages. We propose a first model for CaMEL that uses a massively multilingual corpus to extract case markers in 83 languages based only on a noun phrase chunker and an alignment system. To evaluate CaMEL, we automatically construct a silver standard from UniMorph. The case markers extracted by our model can be used to detect and visualise similarities and differences between the case systems of different languages as well as to annotate fine-grained deep cases in languages in which they are not overtly marked.

1 Introduction

What is a case? Linguistic scholarship has shown that there is an intimate relationship between morphological case marking on the one hand and semantic content on the other (see Blake (1994) and Grimm (2011) for overviews). For example, the Latin case marker *-ibus*[1] (Ablative or Dative Plural) can express the semantic category of location. It has been observed that there is a small number of such semantic categories frequently found cross-linguistically (Fillmore, 1968; Jakobson, 1984), which are variously called *case roles* or *deep cases*. Semiotically, the described situation is complicated by the fact that the relationship between case markers and expressed semantic categories is seldom isomorphic, i.e., there is both *case polysemy* (one case, several meanings) and *case homonymy* or *case syncretism* (several cases, one marker) (Baerman, 2009). As illustrated in Figure 1, the Latin Ablative marker *-ibus* can express the semantic category of instrument besides location (case polysemy), and it is also the marker of the Dative Plural expressing a recipient (case syncretism). In addition, there is *case synonymy* (one case, several markers), which further complicates morphosemiotics; e.g., in Latin, *-is* is an alternative marker of the Ablative Plural.

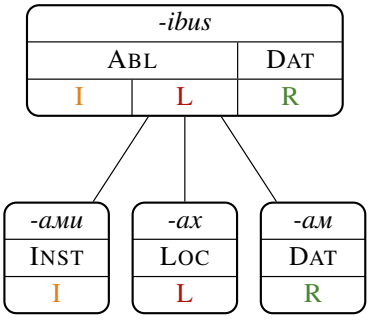

Figure 1: Morpho-semiotic foundation of this study. The Latin case marker *-ibus* is used for both the Ablative (ABL) and the Dative (DAT), which in turn express the three semantic categories of instrument (I), location (L), and recipient (R). This is an example of both case polysemy (one case: ABL, several meanings: I and L) and case syncretism (several cases: ABL and DAT, one marker: *-ibus*). Russian, on the other hand, has an isomorphic relationship between Instrumental (INST), Locative (LOC), and Dative (DAT), the case markers corresponding to them (*-ами*/*-ami*, *-ах*/*-ax*, *-ам*/*-am*), and the expressed semantic categories (I, L, R).

The key idea of this paper is to detect such complex correspondences between case markers and expressed semantic categories in an automated way. Specifically, we build on prior work by Cysouw (2014), who lays the theoretical foundation for our study by showing that deep cases can be induced from cross-linguistic usage patterns of case markers. As opposed to Latin, Russian has separate cases (with separate case markers) for the semantic categories of instrument (*-ами*/*-ami*), location (*-ах*/*-ax*), and recipient (*-ам*/*-am*). Thus, knowing the Russian case marker corresponding to Latin *-ibus* reduces the uncertainty about the expressed

[1]In this paper, we use *italic* when talking about case markers as morphemes in a linguistic context and `monospace` (accompanied by `$` to mark word boundaries) when talking about case markers in the context of our model. Transliterations of Cyrillic examples are given after slashes.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5506 - 5516

case role (Figure 1). This reduction of uncertainty can be particularly helpful in a low-resource setting where other means of analysis are unavailable.

In this work, we rely on the Parallel Bible Corpus (PBC; Mayer and Cysouw, 2014), a massively multilingual corpus, to investigate the relationship between surface cases and their deep meanings cross-linguistically. To put our idea into practice, we require an exhaustive set of case markers as well as a set of parallel noun phrases (NPs) that we can further analyze with respect to deep cases using the set of case markers. Both requirements pose a serious challenge for languages with limited available resources. We therefore introduce **CaMEL** (**Ca**se **M**arker **E**xtraction without **L**abels), a novel and challenging task of finding case markers using only (i) a highly parallel corpus covering many languages, (ii) a noun phrase chunker for English, and (iii) word-level pre-computed alignments across languages.

Our work uses the parallel nature of the data in two ways.

First, we leverage the word-level alignments for the initial step of our pipeline, i.e., the marking of NPs in all languages (even where no noun phrase chunker is available). To do so, we mark NPs in 23 different English versions of the Bible and project these annotations from each English to each non-English version using the word-level alignments, resulting in parallel NPs that express the same semantic content across 83 languages. Based on the projected annotations, we leverage the frequencies of potential case markers inside and outside of NPs as a filter to distinguish case markers from lexical morphemes and other grammatical morphemes typically found outside of NPs.

Second, we leverage the alignments for a fine-grained analysis of the semantic correspondences between case systems of different languages.

We make three main **contributions**.

- We define **CaMEL** (**Ca**se **M**arker **E**xtraction without **L**abels), a new and challenging task with high potential for automated linguistic analysis of cases and their meanings in a multilingual setting.
- We propose a simple method for CaMEL that is efficient, requires no training, and generalises well to low-resource languages.
- We automatically construct a silver standard based on human-annotated data and evaluate our method against it, achieving an F1 of 45%.

To foster future research on CaMEL, we make the silver standard, our code, and the extracted case markers publicly available[2].

2 Related Work

Unsupervised morphology induction has long been a topic of central interest in natural language processing (Yarowsky and Wicentowski, 2000; Goldsmith, 2001; Schone and Jurafsky, 2001; Creutz and Lagus, 2002; Hammarström and Borin, 2011). Recently, unsupervised inflectional paradigm learning has attracted particular interest in the research community (Erdmann et al., 2020; Jin et al., 2020), reflected also by a shared task devoted to the issue (Kann et al., 2020). Our work markedly differs from this line of work in that we are operating on the level of case markers, not full paradigms, and in that we are inducing morphological structure in a massively multilingual setting.

There also have been studies on extracting grammatical information from text by using dependency parsers (Chaudhary et al., 2020; Pratapa et al., 2021) and automatically glossing text (Zhao et al., 2020; Samardžić et al., 2015) as well as compiling full morphological paradigms from it (Moeller et al., 2020). By contrast, our method is independent of such annotation schemata, and it is also simpler as it does not aim at generating full grammatical or morphological descriptions of the languages examined. There has been cross-lingual work in computational morphology before (Snyder and Barzilay, 2008; Cotterell and Heigold, 2017; Malaviya et al., 2018), but not with the objective of inducing inflectional case markers.

Methodologically, our work is most closely related to the SuperPivot model presented by Asgari and Schütze (2017), who investigate the typology of tense in 1,000 languages from the Parallel Bible Corpus (PBC; Mayer and Cysouw, 2014) by projecting tense information from languages that overtly mark it to languages that do not. Based on this, Asgari and Schütze (2017) perform a typological analysis of tense systems in which they use different combinations of tense markers to further divide a single tense in any given language. Our work differs in a number of important ways. First, we do not manually select a feature to investigate

[2] https://github.com/LeonieWeissweiler/CaMEL

but model all features in our chosen sphere of interest (i.e., case) at once. Furthermore, we have access to word-level rather than verse-level alignments and can thus make statements at a more detailed resolution (i.e., about individual NPs). Finally, we extract features not only for a small selection of pivot languages, but even for languages that do not mark case "non-overtly", i.e., in a way that deviates to a large degree from a simple 1–1 mapping (see discussion in §1).

3 Linguistic Background

There is ongoing discussion in linguistic typology about the extent to which syntactic categories are shared and can be compared between the world's languages (see Hartmann et al. (2014) for an overview). While this issue is far from being settled, there is a general consensus that (while not being a language universal) there is a core of semantic categories that are systematically found cross-linguistically, and that are expressed as morphosyntactic case in many languages. Here, we adopt this assumption without any theoretical commitment, drawing upon a minimal set of deep cases detailed in Table 1. The set is loosely based on the classical approach presented by Fillmore (1968).

Going beyond deep cases, Cysouw (2014) envisages a more fine-grained analysis of what is conventionally clustered in a deep case or semantic role. Briefly summarised, the theoretical concept is this: if every language has a slightly different case system, with enough languages it should be possible to divide and cluster NPs at any desired level of granularity, from the conventional case system down to a specific usage of a particular verb in conjunction with only a small set of nouns. For example, the semantic category of location could be further subdivided into specific types of spatial relationships such as 'within', 'over' and 'under'. Taken together, it would then be possible to perform theory-agnostic typological analysis of case-like systems across truly divergent and low-resource languages by simply describing any language's case system in terms of its clustering of very fine-grained semantic roles into larger systems that are overtly marked.

The approach sketched in the last paragraph is not limited to case systems but has been applied to person marking (Cysouw, 2008), the causative/inchoative alternation (Cysouw, 2010), and motion verbs (Wälchli and Cysouw, 2012). The variety of linguistic application areas highlights the potential of developing methods that are much more automated than the work of Cysouw and collaborators. While we stay at the level of traditional deep cases in this paper, we hope to be able to extend our method into the direction of a more general analysis tool in the future.

The remainder of the paper is structured as follows. Section 4 describes our method in detail. Section 5 gives an overview of our results. Finally, Section 6 presents two exploratory analyses.

4 Methodology

4.1 Data

We work with the subset of the PBC (Mayer and Cysouw, 2014) for which the SimAlign alignment algorithm (Jalili Sabet et al., 2020) is available, resulting in 87 languages for our analysis. From the corpus, we only extract those verses that are available in all languages, thus providing for a relatively fair comparison, and remove Malagasy, Georgian, Breton, and Korean, as they have much lower coverage than the other languages. This leaves us with 83 languages and 6,045 verses as our dataset. We also select 23 English versions from the PBC that cover the same 6,045 verses. For each of the 6,045 verses, we then compute $83 \times 23 = 1909$ verse alignments: 83 (for each language) multiplied with 23 (for each English version). In the following, we will describe the components of our pipeline (Figure 2).

4.2 NP Annotation

Because our intermediate goal is to induce complete lists of case markers in all languages we cover, the first step is to restrict the scope of our search to NPs. We hope that this will allow us to retrieve case markers for nouns and adjectives while disregarding verb endings that might otherwise have similar distributional properties. As we are working with 83 languages, most of which are low-resource and lack high-quality noun phrase chunkers, we first identify NPs in English using the spaCy noun phrase chunker (Honnibal et al., 2020) and then project this annotation using the alignments to mark NPs in all other languages. The exception to this are German and Norwegian Bokmål, for which noun phrase chunkers are available directly in spaCy. Because both the spaCy noun phrase chunker and the alignments are prone to error, we make use of 23 distinct English versions

Deep Case	Description	Example
Nominative	The subject of the sentence	He is the Messiah!
Genitive	An entity that possesses another entity	Are you the Judean People's Front?
Recipient	A sentient destination	I gave the gourd to Brian.
Accusative	The direct object of the sentence	Consider the lilies.
Locative	The spatial or temporal position of an entity	They haggle in the market.
Instrumental	The means by which an activity is carried out	The graffiti was written by hand.

Table 1: Descriptions and examples for the deep cases distinguished in this paper, which loosely follow the core deep cases proposed in the classical approach of Fillmore (1968).

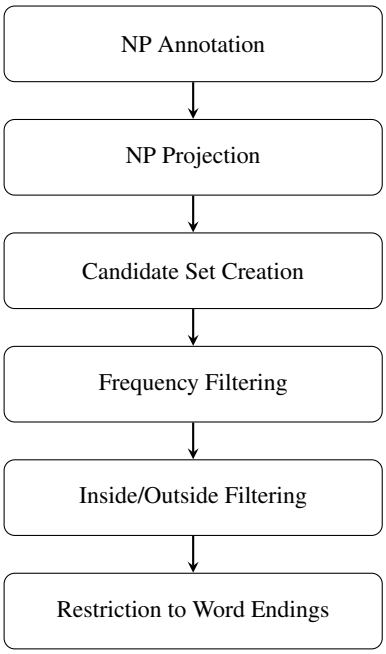

Figure 2: Overview of our pipeline.

of the Bible and mark the NPs in each of them with the goal of lessening the impact of noise.

4.3 NP Projection

We project the NP annotation of a given English version to a second language using the alignments. Specifically, we find the NP in the target language by following the alignments from all words in the English NP while maintaining the word order of the target sentence. We treat each annotated version of the corpus resulting from the different English versions as a separate data source. As an example, Figure 3 shows two English versions and the NP projections for Latin and German. While the alignments, particularly those from English to Latin, are not perfect, they result in complementary errors. The first wrongly aligns the first mention of *pastor bonus*, resulting in only *pastor* being marked as an NP. The second misses the alignment of *life* and *animam*. In these two cases, the other alignment corrects the error.

There are two major results from this process.

First, we obtain the set N of all NPs marked in English, each with all of its translations in the other languages. An example of an entry in this set, taken from Figure 3, would be *the fine shepherd, pastor bonus, der vortreffliche Hirte, ...*, while *the fine shepherd, pastor, der vortreffliche Hirte, ...* would be another, slightly defective, example.

Second, we obtain a pair of multisets, W^l_{in} and W^l_{out}, one for each language l. W^l_{in} (resp. W^l_{out}) is the multiset of all word tokens that appear inside (resp. outside) of NPs of language l. In the following, we will use $M(w)$ to refer to the frequency of word w in the multiset M.

For each language, we want to remove false positives from the word types contained within NPs (which are an artefact of wrong alignments) by using the frequency of each word type inside and outside of NPs.

In principle, this could be done by means of a POS tagger and concentrating on nouns, adjectives, articles, prepositions, and postpositions, but as we do not have access to a reliable POS tagger for most languages covered here, we use the relative frequency information gained from our NP annotations. More specifically, we assign each word type $w \in W^l_{\text{in}} \cup W^l_{\text{out}}$ to I_l (the set of words for language l that are NP-relevant) if $|W^l_{\text{in}}(w)| > |W^l_{\text{out}}(w)|$, and to O_l (the set of words for language l that are not NP-relevant) otherwise. This enhances the robustness of our method against occasional misannotations: for Latin, *ovibus* 'sheep', from our previous example, occurred once outside an NP but 45 times inside and is now an element of I_{Latin}, while *intellegent* 'they understand' occurred once inside an NP but 22 times outside and is therefore an element of O_{Latin}.

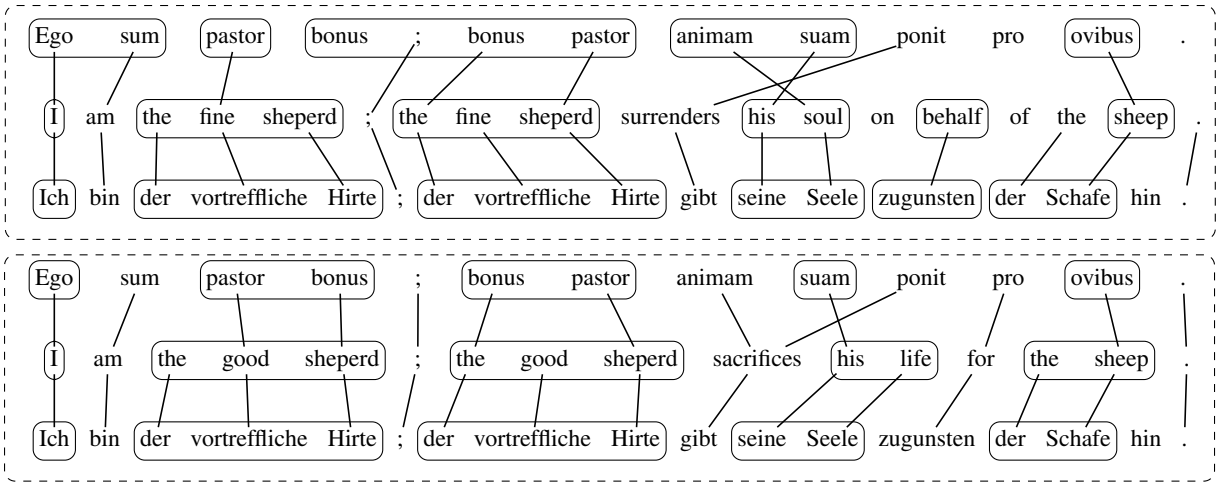

Figure 3: Example of alignments and NP projections (English to Latin and English to German) with two different English versions (top and bottom).

4.4 Candidate Set Creation

From each language, we create a set of candidate case markers $\text{candidates}(w)$ for a word w by collecting all character n-grams of any length from w that are also members of I_l. We explicitly mark the word boundaries with `$` so that n-grams in the middle of words are distinct from those at the edges. For example, candidates extracted from *ovibus* would be `$ovi`, `ibus$`, but also `$ovibus$` and `i`. Our first candidate set is computed as $C_1^l = \bigcup\{\text{candidates}(w) \mid w \in I_l\}$.

4.5 Frequency Filtering

We define $I_l(c)$ as the number of words in I_l that contain the candidate c, and $O_l(c)$ analogously for O_l. As a first step, we filter out all n-grams with a frequency in I_l lower than a threshold θ.[3] This results in $C_2^l = \{c \mid c \in C_1^l, I_l(c) \geq \theta\}$.

4.6 Inside/Outside Frequency Filtering

For this step, we make use of the observation that case is a property of nouns. Hence, a case marker is expected to occur much more frequently within NPs. This will serve to distinguish the case markers from verb inflection markers, which should otherwise have similar distributional properties. To implement this basic idea, for each candidate c in language l, we first construct the contingency table shown in Table 2.

We use the table to test whether a candidate is more or less likely to appear inside NPs by comparing the frequencies of the candidate inside and outside NPs to those of all other candidates. Shown in the cells are the frequencies used for the test for each candidate. The columns correspond to the frequency of the candidate in question versus all other candidates while the rows distinguish the frequencies inside versus outside NPs. We carry out a Fisher's Exact Test (Fisher, 1922) on this table, which gives us a p-value and an odds ratio r. $r < 1$ if the candidate is more likely to occur outside an NP, and $r > 1$ if it is more likely to occur inside. The p-value gives us a confidence score to support this ratio (lower is better). We keep for C_{final}^l only those candidates for which $p < \phi$ and $r > \chi$.[4] For example, `ibus$` makes it past this filter with $p(\texttt{ibus\$}) = 2.869 \cdot 10^{-6}$ and $r(\texttt{ibus\$}) = 1.915$ – it is significant and it occurs inside NPs more often than outside NPs. In contrast, `t$` is discarded as it has $p(\texttt{t\$}) = 3.18 \cdot 10^{-149}$ and $r(\texttt{t\$}) = 0.249$ – it is significant, but it has been found to occur much more likely outside than inside NPs.

	c	$\neg c$
NP	$I_l(c)$	$\sum_{c' \neq c \in C_2^l} I_l(c')$
$\neg$ NP	$O_l(c)$	$\sum_{c' \neq c \in C_2^l} O_l(c')$

Table 2: Contingency table for candidate case marker c in language l for inside/outside filtering. A morphological marker that occurs significantly more often inside NPs than outside of NPs is likely to be a nominal case marker.

4.7 Restriction to Word Endings

Suffixoidal inflection is cross-linguistically more common than prefixoidal and infixoidal inflection (Bauer, 2019). This is also reflected in our dataset,

[3] We set $\theta = 97$ based on grid search.

[4] We set $\phi = 0.08$ and $\chi = 0.34$ based on grid search.

where not a single language has prefixoidal or infixoidal inflection. We hence restrict the set of considered n-grams to ones at the end of words.

5 Evaluation of Retrieved Case Markers

We evaluate our method for case marker extraction without labels using a silver standard.

5.1 Silver Standard

As we are, to the best of our knowledge, the first to introduce this task, we cannot rely on an existing set of gold case markers for each language we cover. As most of the languages included are low-resource, reliable grammatical resources do not always exist, which makes the handcrafting of a gold standard difficult. Therefore, and also to ensure relative comparability, we evaluate against a silver standard automatically created from the UniMorph (Sylak-Glassman 2016, Kirov et al. 2018, McCarthy et al. 2020)[5] dataset. The UniMorph data consists of a list of paradigms, which we first filter by their POS tag, keeping only nouns and adjectives and filtering out verbs and adverbs. An example of a paradigm is given in Table 3. While the Nominative Singular (left column) is included in addition to the inflected forms (middle column), the straightforward approach of extracting the suffixes of the inflected forms is not optimal for every language, as the Nominative Singular form can differ from the root. We therefore proceed as follows.

First, we form a multiset of all inflected forms. In our example, this would result in {*Abflug, Abfluges, Abflug, Abflug, Abflüge, Abflüge, Abflügen, Abflüge*}. Next, we iterate over this multiset, removing one word each time if it occurs only once. This is meant to make the algorithm more robust against outlier words which do not share a common base with the rest of the paradigm. We then extract the longest common prefix for the remaining elements. We build a frequency list of these prefixes, which in our example has only one element, *Abfl*, with a frequency of 3. We take the most frequent element from the frequency list and compare it to the Nominative Singular, *Abflug*. Of these two candidates, we take the longer one. We thereby prioritise precision over recall as roots that are too short quickly result in many different suffixes that are too long, due to the high overall number of paradigms. Finally, we iterate over the inflected forms again, extracting the suffix if the chosen root

[5] https://unimorph.github.io

Nominative Singular	inflected forms		unused information
	base	suffix	
Abflug	Abfl ug		N NOM SG
	Abfl ug	es	N GEN SG
	Abfl ug		N DAT SG
	Abfl ug		N ACC SG
	Abfl üge		N NOM PL
	Abfl üge		N GEN PL
	Abfl ügen		N DAT PL
	Abfl üge		N ACC PL

Table 3: Example of silver standard creation. Marked in orange is the Nominative Singular form, in red the base ("base") as determined by the algorithm, and in green the only suffix ("suffix") that is extracted from this paradigm. Additional, unused information in the UniMorph data is marked in grey.

is a prefix, which in our example yields one new suffix: `es$`, as *Abflüge* and *Abflügen* are not prefixed by *Abflug*. We examine the results for each language and exclude the languages where either basic knowledge of the language or common sense makes it apparent that sets are much too large or too small, resulting in a diverse set of 19 languages to evaluate our methods against. We note that this process automatically excludes adpositions and clitics, which is in line with our focus on suffixoidal inflection (Section 4.6). We make our silver standard publicly available.

5.2 Results

Our results are provided in Table 4. We observe that precision is higher, at times even substantially, than recall for most languages contained in the silver standard. Looking at Table 5 as an example, we can see that low precision is mostly due to retrieved case markers being longer (`ение$`/*enie*) or shorter (`й$`/*j*) than the correct ones. It is one of the main challenges in this task to select the correct length of a case marker from a series of substring candidates. The shorter substrings will automatically be more frequent and often correct, but this is not easily solved by a frequency threshold, which excludes other correct candidates that are naturally less frequent. Additionally, we observe that some recall errors are due to an incorrect length of n-grams in the silver standard (`ьям`/*'jam*), highlighting that this issue also exists in its creation process, and suggesting that our performance might even improve when measured against handcrafted data.

Language	P	R	F1
Albanian	.74	.47	.58
Belarusian	.43	.41	.42
Bengali	.50	.40	.44
Czech	.50	.58	.54
German	.54	.47	.50
Greek	.67	.19	.30
Icelandic	.83	.31	.45
Indonesian	.31	.42	.36
Irish	.42	.30	.35
Latin	.65	.56	.60
Lithuanian	.18	.38	.24
Nynorsk	.79	.48	.59
Bokmål	.67	.45	.54
Polish	.52	.33	.40
Russian	.54	.54	.54
Slovenian	.41	.28	.33
Swedish	.68	.25	.36
Ukrainian	.45	.48	.47
Average	.54	.41	.45

Table 4: Precision (P), Recall (R) and F1 on the task of case marker extraction without labels for languages contained in our silver standard. Nynorsk and Bokmål are two varieties of Norwegian.

5.3 Ablation Study

We conduct an ablation study to assess the effects of the different pipeline components.

5.3.1 Evaluating NP Projection

In order to evaluate how well our method of projecting NP annotation using alignments to languages without an available NP chunker (see Section 4.3) works, we evaluate it against the monolingual spaCy chunkers for Norwegian Bokmål and German, which are the only available languages besides English. We do not directly compare annotated spans but instead their influence on our method as we have intentionally designed our pipeline to be robust to some noise. As I_l, the set of words considered to be NP-relevant, is the essential output of the annotation projection, we compare two versions, the set as a result of direct NP chunking and the set as a result of our annotation procedure. Taking the former as the ground truth for evaluating the latter (assuming that the directly chunked set has superior quality), we observe an F1 of 88.5 % for German and 67.8 % for Norwegian Bokmål. While these numbers seem low at first, the fact that our overall F1 on Norwegian Bokmål (.54, see Table 4) is better than on German (.50) indicates that the later elements of the pipeline are to a certain extent robust against misclassification of NPs.

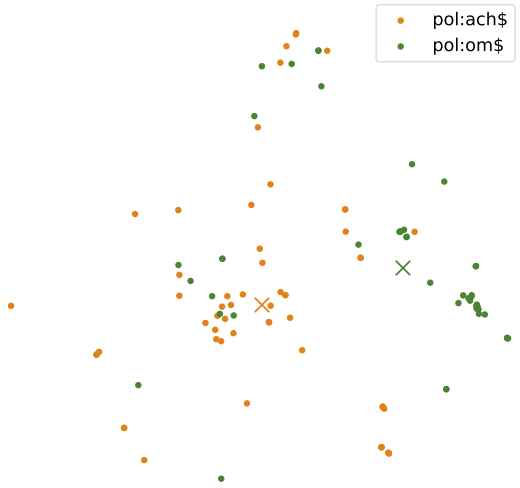

Figure 4: t-SNE plot of the contextual distribution of the Latin case marker *-ibus* and the Polish case markers *-ach* and *-om*. Outliers omitted. The plot shows NPs who in Latin are marked with the case marker `ibus$` and in Polish either with `ach$` (orange) or `om$` (green). Centroids are marked with an X. The plot shows that the Polish case markers exhibit a more fine-grained representation of the underlying semantic categories, which makes it possible to disambiguate the homonymous Latin case marker.

5.3.2 Ablating Pipeline Components

We report the average Precison, Recall, and F1 across all languages in our silver standard without individual filtering components in Table 6. Simple frequency filtering (see "$\neg\theta$"), excluding n-grams within words (see "middle") and at the beginning of words (see "beginning") are all necessary for good performance. Inside/outside filtering based on p-value is the most important component of the pipeline (see "$\neg\phi$"). Surprisingly, inside/outside filtering based on odds ratio has almost no effect.

6 Exploratory Analyses

We can use our automatically extracted case markers, in combination with the parallel NPs that are extracted as part of the pipeline, for innovative linguistic analyses. We present two examples in this section.

6.1 Marking of Deep Cases

First, we demonstrate how, given a parallel NP, the case markers can be used to determine its deep

Intersection	Algorithm Only	Silver Standard Only
у, я, ом, ого, о, в, ой, и, ми, ам, ей, ю, ы, ов, ых, а, м, х, ами	ий, ные, ое, ение, ии, го, ый, ка, ые, к, ки, ия, ние, й, ния, ие	ыми, ах, ев, ьям, ому, ья, н, ьях, ями, ям, е, ях, ьев, ем, ым, ьями
u, ja, om, ogo, o, v, oj, i, mi, am, ej, ju, y, ov, yx, a, m, x, ami	*ij, nye, oe, enie, ii, go, yj, ka, ye, k, ki, ija, nie, j, nija, ie*	*ymi, ax, ev, 'jam, omu, 'ja, n, 'jax, jami, jam, e, jax, 'ev, em, ym, 'jami*

Table 5: The output of our algorithm for Russian compared to the silver standard. We show suffixes that occur in the intersection of algorithm output and silver standard ("Intersection"), those that occur only in the algorithm output ("Algorithm Only") and those that occur only in the silver standard ("Silver Standard Only"). To allow for a clear and concise presentation, the table does not observe the convention of using $ for boundaries.

	ablation	P	R	F1
our method (Table 4)		.54	.41	.45
	$\neg\theta$	.11	.59	.16
	$\neg\phi$	.00	.00	.00
	$\neg\chi$	.53	.41	.44
	middle	.11	.41	.17
	beginning	.33	.41	.35

Table 6: Precision (P), Recall (R) and F1 averaged over all languages on the task of case marker extraction without labels when each step of our pipeline is ablated. $\neg\theta$: no Frequency Filtering; $\neg\phi$: no Inside/Outside Filterung based on p-value; $\neg\chi$: no Inside/Outside Filtering based on odds ratio; middle: include middle of the word; beginning: include beginning of the word.

case. We return to N (see Section 4.3), our set of parallel NPs extracted from the PBC, and for a selected subset of languages, group them by their combination of case markers. The basic idea is to infer an NP's (potentially very fine-grained) deep case by representing it as its combination of case markers across languages.

For example, we can disambiguate the Latin case marker *-ibus* by looking at the different groups the NPs containing it form with Russian case markers. Recall that *-ibus* can express location, instrument, and recipient and that Russian expresses these categories by separate case markers: *-ax*/*-ax* for location, *-ами*/*-ami* for instrument, and *-ам*/*-am* for recipient (see Figure 1) – all three of which have been retrieved by our method. Given a Latin NP marked by the ending *-ibus*, the parallel NP in Russian can help us determine its deep case. Thus, for domibus, дворцах/*dvorcax* shows that the semantic category is location, i.e., 'in the houses'. For operibus bonis, добрыми делами/*dobrymi delami* shows that the semantic category is instrument, i.e., 'through the good deeds'. Finally, for patribus, предкам/*predkam* shows that the semantic category is a recipient, i.e., 'for/to the parents'.

6.2 Similarities between Case Markers

We also demonstrate how we can use their distributional similarities over NPs to show how case markers that are similar in this respect correspond to similar combinations of deep cases. We first generate an NP-word cooccurrence matrix over the NP vocabulary of all languages in which each row, corresponding to an inflected word firm w in language l, indicates which NPs (corresponding to columns) cooccur with w. in the parallel data. We then reduce the dimensionality of the matrix by means of t-SNE (Van der Maaten and Hinton, 2008), allowing us to inspect systematic patterns with respect to the "contexts" in which certain case markers occur (where "context" refers to words the case marker is aligned to in other languages, not words the case marker coccurs with in its own language). In a semiotic situation like the one shown in Figure 1, this setup allows us to examine how the semantic region expressed by a certain homonymous case marker in one language is split into more fine-grained regions in another language that distinguishes the semantic categories that are lumped together by the case marker (and which, if they are at the right level of abstraction, can correspond to deep cases).

Figure 4 shows this scenario for the Latin Ablative marker *-ibus*. It corresponds to two distinct case markers in Polish, *-ach* (LOC) and *-om* (DAT). The figure shows that the region occupied by Latin *-ibus* splits into two distinct clusters in Polish, allowing us to visually determine which underlying case is expressed by the homonymous suffix *-ibus*.

This underscores the exploratory potential of our approach.

7 Conclusion and Future Work

We have introduced the new and challenging task of Case Marker Extraction without Labels (CaMEL) and presented a simple and efficient method that leverages cross-lingual alignments and achieves an F1 of 45% on 19 languages. We introduce an automatically created silver standard to conduct our evaluation. We have further demonstrated two ways in which our retrieved case markers can be used for linguistic analysis.

We see two potential avenues for future work. The first is the further improvement of case marker extraction. The main problem to tackle here is that of small sets of overlapping substrings of which only one is the correct marker, and developing some further measures by which they can be distinguished. Furthermore, it would be useful to find data from more low-resource languages and languages that have typological properties different from the extensively studied large language families (Indo-European, Turkic, Sino-Tibetan etc.). We could then verify that our method performs well across languages and attempt to expand our silver standard to more languages while still ensuring quality. The second area is that of further automating the analysis of deep case and case syncretism. Ideally, we would develop a method that can distinguish the different possible reasons for divergent case marking in languages, with the eventual goal of creating a comprehensive overview of case and declension systems for a large number of languages.

Acknowledgements

This work was funded by the European Research Council (#740516). The second author was also supported by the German Academic Scholarship Foundation and the Arts and Humanities Research Council. The third author was also supported by the German Federal Ministry of Education and Research (BMBF, Grant No. 01IS18036A). We thank the reviewers for their extremely helpful comments.

References

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark. Association for Computational Linguistics.

Matthew Baerman. 2009. Case syncretism. In Andrej Malchukov and Andrew Spencer, editors, *The Oxford handbook of case*, pages 219–230. Oxford University Press, Oxford, UK.

Laurie Bauer. 2019. *Rethinking morphology*. Edinburgh University Press, Edinburgh, UK.

Barry J. Blake. 1994. *Case*. Cambridge University Press, Cambridge, UK.

Aditi Chaudhary, Antonios Anastasopoulos, Adithya Pratapa, David R. Mortensen, Zaid Sheikh, Yulia Tsvetkov, and Graham Neubig. 2020. Automatic extraction of rules governing morphological agreement. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5212–5236, Online. Association for Computational Linguistics.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. Association for Computational Linguistics.

Michael Cysouw. 2008. Building semantic maps: The case of person marking. In *New challenges in typology*, pages 225–248. De Gruyter Mouton.

Michael Cysouw. 2010. Semantic maps as metrics on meaning. *Linguistic Discovery*, 8(1):70–95.

Michael Cysouw. 2014. Inducing semantic roles. *Perspectives on semantic roles*, 23:68.

Alexander Erdmann, Micha Elsner, Shijie Wu, Ryan Cotterell, and Nizar Habash. 2020. The paradigm discovery problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7778–7790, Online. Association for Computational Linguistics.

Charles J. Fillmore. 1968. The case for the case. In Emmon Bach and Robert T. Harms, editors, *Universals in linguistic theory*, pages 1–88. Holt, Rinehart & Winston, New York, NY.

Ronald A Fisher. 1922. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

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

Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309–350.

Iren Hartmann, Martin Haspelmath, and Michael Cysouw. 2014. Identifying semantic role clusters and alignment types via microrole coexpression tendencies. *Studies in Language*, 38(3):463–484.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Roman Jakobson. 1984. Contribution to the general theory of case: General meanings of the Russian cases. In Linda R. Waugh and Morris Halle, editors, *Russian and Slavic grammar: Studies 1931-1981*, pages 59–103. De Gruyter, Berlin.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Huiming Jin, Liwei Cai, Yihui Peng, Chen Xia, Arya McCarthy, and Katharina Kann. 2020. Unsupervised morphological paradigm completion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6696–6707, Online. Association for Computational Linguistics.

Katharina Kann, Arya D. McCarthy, Garrett Nicolai, and Mans Hulden. 2020. The SIGMORPHON 2020 shared task on unsupervised morphological paradigm completion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 51–62, Online. Association for Computational Linguistics.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Chaitanya Malaviya, Matthew R. Gormley, and Graham Neubig. 2018. Neural factor graph models for cross-lingual morphological tagging. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2653–2663, Melbourne, Australia. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Sarah Moeller, Ling Liu, Changbing Yang, Katharina Kann, and Mans Hulden. 2020. IGT2P: From interlinear glossed texts to paradigms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5251–5262, Online. Association for Computational Linguistics.

Adithya Pratapa, Antonios Anastasopoulos, Shruti Rijhwani, Aditi Chaudhary, David R. Mortensen, Graham Neubig, and Yulia Tsvetkov. 2021. Evaluating the morphosyntactic well-formedness of generated texts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7131–7150, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tanja Samardžić, Robert Schikowski, and Sabine Stoll. 2015. Automatic interlinear glossing as two-level sequence classification. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 68–72, Beijing, China. Association for Computational Linguistics.

Patrick Schone and Daniel Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Benjamin Snyder and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745, Columbus, Ohio. Association for Computational Linguistics.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (unimorph schema). *Johns Hopkins University*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Bernhard Wälchli and Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics*, 50(3):671–710.

David Yarowsky and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 207–216, Hong Kong. Association for Computational Linguistics.

Xingyuan Zhao, Satoru Ozaki, Antonios Anastasopoulos, Graham Neubig, and Lori Levin. 2020. Automatic interlinear glossing for under-resourced languages leveraging translations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5397–5408, Barcelona, Spain (Online). International Committee on Computational Linguistics.