Comparative Concepts or Descriptive Categories: a UD Case study

Matthieu Pierre Boyer
Lattice
DI ENS
Paris, France
matthieu.boyer@ens.fr

Mathieu Dehouck
Lattice, CNRS, ENS-PSL, USN
mathieu.dehouck@cnrs.fr

Abstract

In this paper, we present a series of methods used to quantify the soundness of using the same names to annotate cases in different languages. We follow the idea described by Martin Haspelmath that descriptive categories and comparative concepts are different objects and we look at the necessary simplification taken by the Universal Dependencies project. We thus compare cases in closely related languages as belonging to commensurable descriptive categories. Then we look at the corresponding underlying comparative concepts. We finally looked at the possibility of assigning cases to adpositions.

1 Introduction

> There is a fundamental distinction between language-particular categories of languages (which descriptive linguists must describe by descriptive categories of their descriptions) and comparative concepts (which comparative linguists may use to compare languages).
>
> *Martin Haspelmath* in (Haspelmath, 2018)

Language description and language comparison are two intertwined yet distinct endeavours. Language description is often done in a language different from the one being described (many grammars have been written in English, French, Russian, Spanish and Portuguese for example) and often uses a conventionalised descriptive meta-language associated with a given descriptive school. Language comparison relies on the previous step of language description as it main data source but also needs a common meta-language to name the various phenomena under study.

In his paper, Haspelmath (2018) warns us against the confusion of the different meta-languages (the descriptive languages used in each individual description and the common comparative meta-language). He advocates for a careful choice of terms when describing similar categories across multiple languages, even when the similarities compel us to use the same term. That is, one should avoid using a single term to describe two categories from two different languages. Even more so, when this term is also used as a comparative concept which then further increases the risk of cross-meta-language confusion.

With all its qualities, the Universal Dependencies (UD) project (Zeman et al., 2024) puts itself exactly in this somewhat uncomfortable situation. One of the main aims of the project is to foster linguistic typological research, and thus it proposes a common annotation scheme for creating treebanks for all natural languages (de Marneffe et al., 2021). Figure 1 depicts the dependency tree of a Turkish sentence as an example. While the scheme has means to accommodate language specific phenomena, its core is language agnostic and treebank creators are compelled to reuse previously defined language specific extensions when annotating similar structures in new languages as a mean to increase the overall consistency and comparability of the corpora. In the dependency tree, the labels of the edge going out of a node is called its *dependency relation* and the target of the edge it the *governor* of the node. However, the annotation also needs to be sound from the point of view of each annotated language (see points 1 and 2 of the presentation page at https://universaldependencies.org/introduction.html). Each individual treebank can thus be seen as a kind of description

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 55–65
March 3-4, 2025 ©2025 University of Tartu Library

of its language. Indeed, that is exactly what Herrera et al. (2024) do in their work, where they use sparse representation methods to try to extract a grammar sketch for a language from its annotated treebank. In UD, the same terms are thus used both as comparative concepts and as descriptive categories for all the languages that express that category.

In this study, we investigate the descriptive-comparative confusion arising from UD's annotation scheme at the morphosyntactic level. We especially focused on the category of case and its different realisations across several languages with the following question in mind: Do cases sharing their name have the same value across different languages? The main reason to focus on the case category, is that it has both strongly syntactic and strongly semantic values. For example, in languages with a case marking the subject of both transitive and intransitive verbs, this case is usually called NOMINATIVE[1] based on its syntactic properties. If the same language has another case marking the "together with" relation, it will usually be called COMITATIVE on semantic ground.

This study should provide insight on the extent to which one can transfer information about a feature from a language to another simply by reusing the same name (using the same descriptive category). In the end, it could help improve cross-lingual learning scenarios where we want to use as much information from other languages as we can, even at the morphological and syntactic levels.

This paper is organised as follows. Section 2 gives an overview of UD's guidelines on case annotation and how these are realised in practice. Section 3 describes how we assign representations to cases. Section 4 looks at the similarity between cases from different languages as if they were descriptive categories. Section 5 then turns to looking at cases as comparative concepts applied to each individual treebank. Section 6 takes an in between look directly at the cases from all the treebanks. Section 7 investigates the possibility of assigning cases directly to adpositions. Eventually, Section 8 concludes this paper.

1.1 Theoretical Note

In this work, we decided to question the relevance of using the same name to refer to cases in different languages. This assumes the existence of a commensurable case category in each language of interest. There is however no reason to take it for granted.

We decided to take a very pragmatic stance. Universal Dependencies (and indeed, many linguists) assumes a commensurable case category existing across languages. So, we acknowledge this choice. We neither question the existence of a case category in different languages, nor do we question the number of values displayed by said category in each language of interest. We question the relevance of the names given to the different values in different languages.

2 The `Case` Feature across Treebanks

While realising this study, we stumbled upon a number of incongruities in the way the different corpus use the `Case` feature.

There are essentially three ways the feature `Case` is used in the UD treebanks. The first and by far the most common use is to annotate inflected forms of nouns, pronouns and proper nouns in languages where these words inflect according to their role in a clause, as well as determiners, adjectives and participles in languages where they inflect to match the case of their governor.

The second use that is documented in UD's guidelines[2], is to annotate adpositions with the case they give to their nominal phrase, especially so in languages without over case marking on nouns. This annotation principle indicates that UD leans more toward the application of comparative concepts to individual languages. Indeed, if a language does not use the case category, then the "case" represented by an adposition can only be inferred either by comparing its distribution to the distribution of actual cases in languages that possess that category, or by applying formal comparative definitions.

However, this is not always how this feature is used, as in Czech CLTT treebank (Kríž and Hladká, 2018) for example, adpositions are annotated with the `Case` feature and their value always match that of their governing noun. This is all the

[1] In this paper, we use faces to distinguish between DESCRIPTIVE CATEGORIES, COMPARATIVE CONCEPTS and UD's `annotation scheme`.

[2] See the page of the `Case` feature: https://universaldependencies.org/u/feat/Case.html.

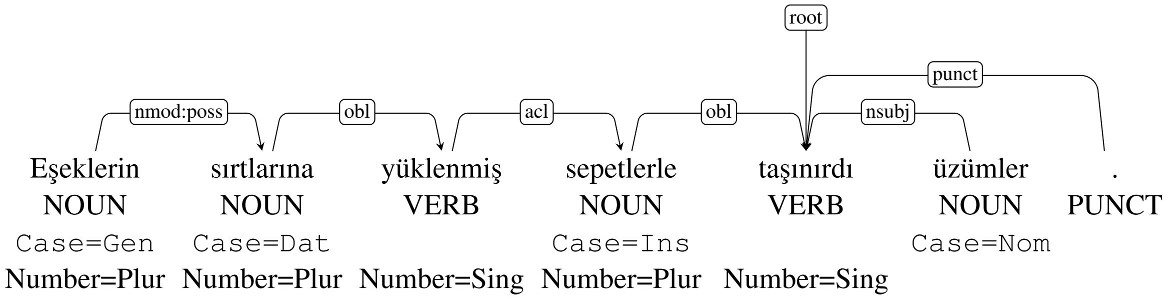

root
punct
nmod:poss
obl
acl
obl
nsubj
Eşeklerin NOUN Case=Gen Number=Plur
sırtlarına NOUN Case=Dat Number=Plur
yüklenmiş VERB Number=Sing
sepetlerle NOUN Case=Ins Number=Plur
taşınırdı VERB Number=Sing
üzümler NOUN Case=Nom
. PUNCT

Figure 1: Representation of the dependency graph of the Turkish sentence ”Eşeklerin sırtlarına yüklenmiş sepetlerle taşınırdı üzümler.” from UD's Turkish BOUN corpus, meaning ”Grapes were carried in baskets loaded on donkeys' backs.”

more surprising that Czech adpositions are invariable and can license several case values.

This indeed points to another problem with case annotation on adpositions. Like languages exhibiting case syncretism[3], adpositions can in principle also be used to mark different syntactic and semantic roles. It becomes then even less clear how one should proceed in assigning cases to adposition.

The third and most divergent use of the `Case` feature can be seen in the Persian Seraji treebank. In this treebank, we only find three case values : `Case=Loc`, `Case=Tem` and `Case=Voc`. The first two values are exclusively used to annotate adverbs of place and adverbs of time respectively. The third value is used to annotate an interjection used to create vocative noun phrase.

3 Case Representation

In order to compare cases from different languages, we need to find a shared representation that should be as language agnostic as possible. We decided to use the syntactic profile of a case defined as the probability distribution[4] over the dependency relations to its governors. This choice is both theoretical since core cases are usually defined in terms of syntactic relations to the other constituents of a sentence, and practical since UD treebanks are annotated with dependency labels.

In order to make the representations even more language agnostic, we decided to ignore relation sub-types since they are not consistently used across languages and corpora. So, both `flat:foreign` and `flat:name` are counted as `flat`.

We give two representations to each case in a language. The first is the empirical probability distribution of the relation of a word displaying that case to its governor.

However, there are several mechanisms underlying case assignment, and not all are as informative. For example, when determiners inflect for case, they usually inherit their value from their head noun, which therefore does not teach us much about that case since a determiner can in principle take any case that way. Similarly, it would artificially separate cases from languages with articles (a high proportion of `det` relations) from those of languages without.

Furthermore, as mentioned in the previous section, UD also allows annotation of the `Case` feature on adpositions, which is quite different from the way cases are generally assigned to nouns. For all these reasons, we thus decided to have a part-of-speech based representation too.

The second representation is thus the syntactic profile of the nouns (`NOUN`) which bear the said case. This gets rid of less informative dependency relations such as `case`, `amod` or `det` and we further decided to ignore the `conj` relation for similar reasons.

The relation distributions are computed from the concatenation of the three parts (train, dev and test) of each treebank from UD version 2.14 (Zeman et al., 2024), except when precised otherwise.

[3] A given word form can be ambiguous as to its morphological features. For example, the Latin form *rosae* can be either a genitive or dative singular or a nominative or vocative plural.

[4] This may be better thought of as normalized frequency distributions, since the case of a word is not a random variable but rather the result of its use in context. But mathematically, normalized frequencies can be viewed as probability distributions.

4 Sharing Descriptive Categories

With our case representations, we first look at cases used in different treebanks as representing the values of a descriptive category. We want to know how relevant is to apply the same name to values of a similar category in different languages.

First, we compare case labels from two closely related languages, namely Czech and Russian[5]. To do so we compute the euclidean distance between each case in the first language and each case in the second language. Then, we generate a 1-nearest neighbour graph assuming the neighbours of a node must come from the other language. This gives us an idea of the way cases could be mapped in a transfer learning setting for example.

Figure 2 represents the 1-nearest neighbour graph of Czech and Russian cases when representations are computed over all the words marked for case. We see that the Czech and Russian NOMINATIVES are each other's nearest neighbour and such is the case for the two genitives. However, for the other cases, the picture is less clear. This is likely due to the fact that when we compute the representations using all the parts-of-speech at once, we confuse the different types of case assignment.

Figure 3 which represents the 1-nearest neighbour graph of Czech and Russian cases when representations are computed only on nouns, is clearer. On top of the NOMINATIVES and GENITIVES, the ACCUSATIVES and INSTRUMENTALS are also each other's nearest neighbours. Only the DATIVES, LOCATIVES and Russian PARTITIVE are still entangled. Looking directly at the data, we realize that the `iobj` relation is never used in the Czech CLLT corpus. The increased probability of seeing a noun in the DATIVE descending from an `obl` relation makes the Czech DATIVE more distinct from the Russian DATIVE and the Czech LOCATIVE is.

The distance matrices for these two graphs can be found in the appendix, along with distance matrices for Czech - Turkish. In the latter, we might for example see that the `equative` behaves erratically on nouns, but that simply comes from the fact that only one noun is annotated with `equative` in the Turkish BOUN corpus.

Note that not all pairs of languages are as well behaved as Czech and Russian, as we shall see in Section 6.

Case	Description	DepRel
NOMINATIVE	Subject of a clause.	`nsubj`
ACCUSATIVE	Direct object of transitive verbs.	`obj`
ABSOLUTIVE	Subject of intransitive verbs and object of transitive verbs.	`nsubj` `obj`
ERGATIVE	Subject of transitive verbs.	`nsubj`
GENITIVE	Noun complement, typically possessor.	`nmod`
DATIVE	Indirect object of verbs, typically recipient of giving verbs.	`iobj`

Table 1: Ideal description of a few cases and corresponding UD's dependency relations.

5 Applying Comparative Concepts

In the previous section we have compared cases from two languages as if they were from a commensurable descriptive category. In this section, we take the other view that Universal Dependencies defines comparative concepts and that the various treebanks are annotated with these concepts. This means that each case has a language agnostic definition and that it is then applied to each language accordingly. Here, the data used is only from the dev part of the treebanks, for computational time reasons.

Since we do not have language agnostic mathematical representations of the various grammatical cases used in UD's annotations, we need to extract them from the available annotated corpora. Since a case profile depends not only on the choice of a language, but also on the sentences in the corpus (replacing a few sentences will generally slightly affect the frequency distribution), we model each comparative case with a random variable taking values from the probability distributions (or normalized frequency distributions) over the set of dependency relations to a word's governor.

Formally, let c be a case, d a dependency relation and $\mathcal{T}$ a treebank. We note $f_{\mathcal{T}}(c, d)$ the frequency at which a word inflected in case c is attached to its governor via a relation of type d in corpus $\mathcal{T}$. Let $\pi_{\mathcal{T}}(c, d) = f_{\mathcal{T}}(c, d) / \sum_{d'} f_{\mathcal{T}}(c, d')$ be the corresponding probability, and $\pi_{\mathcal{T}}(c, \cdot)$ the

[5] We tried a number of pairs and decided to just present Czech and Russian for space reason.

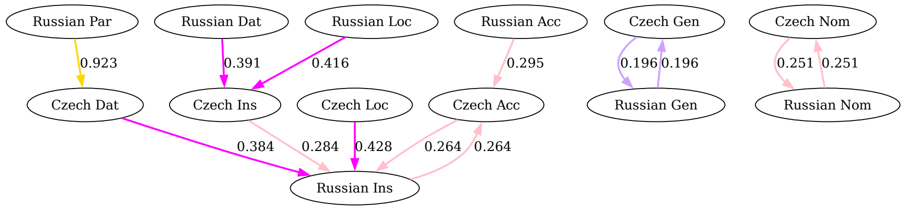
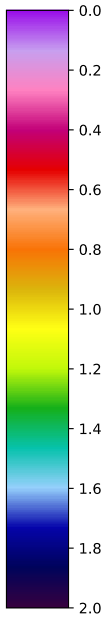


Figure 2: Nearest neighbour graph for Czech CLTT and Russian GSD case profiles. The corresponding distance matrices are Tables 5, 7 and 9 in the appendix.



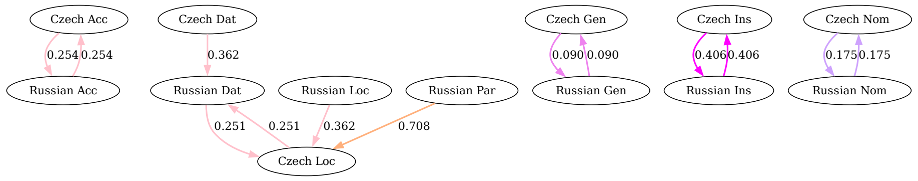


Figure 3: Nearest neighbour graph Czech CLTT and Russian GSD case profiles of nouns. The corresponding distance matrices are Tables 6, 8 and 10 in the appendix.

corresponding probability distribution. We model the case c as a random variable over the probability distributions $\pi_{\cdot}(c, \cdot)$.

We know each of these random variables through a number of realisations: the vector representations of the considered case across all corpora where it is present (which are exactly the probability distributions representations $\pi_{\mathcal{T}}(c, \cdot)$ for each corpus $\mathcal{T}$). That is, this random variable maps a language/corpus to a probability distribution over the dependency relations reaching words marked with that case.

Then, to compute the profile of the comparative cases, we compute the expected value of the random variables associated to each case. Since the values of the random variables are distributions we also compute the barycenter of the realisations of each variable for the Wasserstein 1-distance (or Earth Mover's Distance). We will denote the latter by *Wasserstein barycenter*.

Table 2 gives the representations of the expected distributions of a few selected comparative cases. The representations are mostly aligned with our expectations. But we can still notice a few interesting facts. The ERGATIVE is much more strongly associated with being a subject than the NOMINATIVE is. There may be a few different reasons to that. First, some language like Turkish use the nominative/accusative distinction also to mark a definite/indefinite distinction on the object, with the accusative being kept for definite objects. Another possibility is that when a language has case marking but does not make distinction between subjects and objects such as Irish, it is by default assumed to be nominative-accusative, with the nominative assuming both syntactic roles[6].

Another interesting fact is that the DATIVE's main role is not that of indirect object but rather of oblique. This comes from the strong limitations that UD imposes on the use of the `iobj` relation. But still, DATIVE is virtually the only case to assume that role.

However, while this representation allows us to distinguish many cases syntactically, it doesn't allow to distinguish all cases. More specifically, some cases work in the same syntactic constructions and thus are mostly distinguished through their semantic properties. For example, the Finnish ELLATIVE and ILLATIVE are used to signify that a movement respectively comes from a place or into a place. In the sentence *"I went into his house"*, *house* would be in illative in finnish, while in *"I come back from his house"*, *house* would be in ellative.

This is exactly what we see for non-core cases. LOCATIVE, INSTRUMENTAL and ABLATIVE have very similar profiles, essentially distributed between oblique complements of verbs and nominal

[6] In the eventuality that it would be considered an ergative-absolutive language, the default case would likely be called absolutive rather than ergative anyway.

Case	Average	iobj	nmod	nsubj	obj	obl
ABS	Uniform	0.1	3.3	27.2	36.7	22.4
	Wasserstein	0.0	1.6	28.6	52.2	11.2
ERG	Uniform	0.0	0.7	92.4	0.5	5.9
	Wasserstein	0.0	0.5	97.6	1.4	0.3
NOM	Uniform	0.1	8.0	55.6	7.4	5.0
	Wasserstein	0.0	4.9	65.4	9.3	3.8
ACC	Uniform	0.6	7.8	3.8	62.5	20.5
	Wasserstein	0.0	7.2	1.9	57.6	25.9
GEN	Uniform	0.9	67.4	3.9	5.6	14.9
	Wasserstein	0.0	72.9	3.1	4.5	17.9
DAT	Uniform	14.4	14.9	1.9	0.0	57.2
	Wasserstein	19.0	16.4	0.5	0.0	60.5
LOC	Uniform	0.0	16.6	0.9	1.7	69.6
	Wasserstein	0.0	18.8	0.0	0.0	76.2
INS	Uniform	0.0	17.2	1.4	0.0	66.0
	Wasserstein	0.0	21.3	0.0	0.0	73.8
ABL	Uniform	0.0	16.5	1.3	1.0	70.0
	Wasserstein	0.0	17.2	0.0	0.0	78.5

Table 2: Distributions of the most representative dependency relations for a few cases as computed on nouns. Uniform corresponds to the average profile assuming uniform weighting of each corpus profile. Wasserstein corresponds to barycenters computed with the Wasserstein metric taking into consideration that case profiles are not any vector, but actual probability distributions.

modifiers or nouns.

To check the representativeness of a comparative case P of its realisations across treebanks, we compute also compute its energy E.

$$P = \arg\min_{\mu} E\left(\mu, (\rho_i)_{i\in[\![1,n]\!]}\right) \tag{1}$$

$$E\left(\mu, (\rho_i)_{i\in[\![1,n]\!]}\right) = \frac{1}{n}\sum_{i=1}^{n} d(\mu, \rho_i) \tag{2}$$

The energies associated to the two barycenters are of the same magnitude, with the Wasserstein barycenter being more exacerbated as can be seen in Figure 4 for the ACCUSATIVE case. Here, the ρ_i are ℓ^1-normalized vectors representing cases, and d is the metric used to define the geometry of the space (here, we use the ℓ^2-metric and the Wasserstein 1-distance).

The x-axis represents the different dependency relations leading to nouns in the accusative, the exact list is given in the appendix for convenience. It represents in red the uniform mean of distributions (the expectancy of the variable), in yellow the barycenter of the distributions associated to the Wasserstein 1-distance and in purple the (unnormalized for graphical purposes) apparition frequency.

We can notably see that for uniform mean some relations are represented because very present in a few languages while this is not the case for the Wasserstein barycenter, which is more centered on the dependency relations present in a lot of languages.

6 Case Clustering

In this section we apply data visualisation techniques as a mean to look at the general landscape of case across languages. This is a way to explore similarity between cases for many languages at once and without assuming a prototypical representation for each case.

From a practical annotation perspective, this

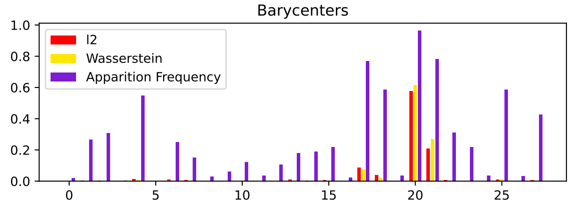


Figure 4: Representation of the uniform barycenter in red and the Wasserstein barycenter in yellow for the comparative ACCUSATIVE case. In purple is represented the proportion of treebanks that associate a given dependency relation with nouns in the accusative from the set of treebanks that inflect their nouns for case.

is interesting since it is more likely too capture the underlying structure of UD's annotations. Indeed, UD's guidelines are sometimes underspecified, which is expected from an annotation scheme whose aim is to be applicable to as many languages as possible. Not all use cases and language specific phenomena will have been thought of during the creation of the guidelines. Therefore, when annotators stumble upon a new structure that does not lend itself to a straightforward analysis, they will both turn to the guidelines and to other treebanks in order to see how similar phenomena might have been annotated in other languages.

We first used a *t-SNE* analysis (van der Maaten and Hinton, 2008) with the hope of seeing well defined clusters. However, plotting all the cases at once proved unmanageable and so we resorted to visualising only a pair of cases each time.

The algorithm consists in looking at the probability distribution generated by the high dimensional vectors[7] representing each instance of the cases and generating a distribution over pairs of those vectors in a way that pairs of *close* vectors are assigned higher probabilities. Then *t-SNE* defines a probability distribution on pairs of 2D points that minimizes the Kullback-Leibler divergence between the two distributions.

Figures 5 and 6 represents the *t-SNE* applied to all the NOMINATIVES and GENITIVES using either the profiles computed on all the words, or just on nouns. It seems that the two cases make for clusters, in the sense they can be grouped along distinct directions. While this is not enough for us to have a classification algorithm, it hints towards possible ways to visualise the difference between cases.

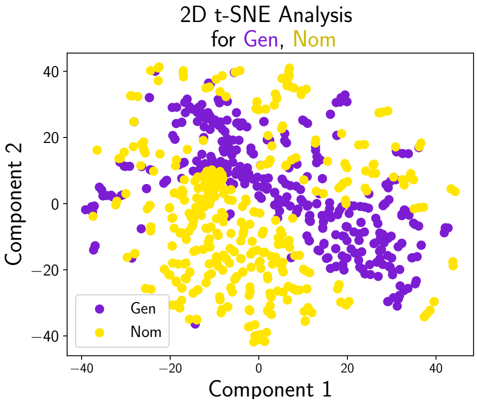


Figure 5: Representation of 2D *t-SNE* analysis of GENITIVE and NOMINATIVE profiles gathered on all the words marked for these cases.

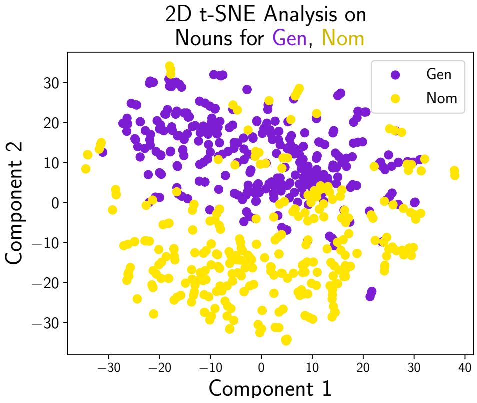


Figure 6: Representation of 2D *t-SNE* analysis of GENITIVE and NOMINATIVE profiles gathered only on the nouns inflected for these cases.

To confirm this hunch we tried to use *ToMATo* (Chazal et al., 2011), a persistence based clustering algorithm, which uses sub-level sets of a function to design a persistence diagram and derive clusters. The implementation that was used comes from Maria et al. (2014). The idea behind *ToMATo* is to compute the density at each point in the representation space and to cluster points using geodesics: every point above a certain elevation and inside the same geodesic belongs in the same cluster (the same hill) and every point below is ignored.

By repeating the process for different elevations[8] we can see clusters appear and merge.

[7] Here the vectors are normalized for the ℓ^1-norm, but we do not consider them as probability distributions

[8] *ToMATo* considers the evolution of the topology of superlevel-sets for α of the density function as α decreases and especially their path-connectivity (or 0-persistence in homological terms).

When two clusters merge, the one with the highest elevation absorbs the other and we say that the lowest one dies. One can then represent on a diagram the birth and death time of each cluster. This is depicted in Figure 7 for GENITIVES and NOMINATIVES. The closer a cluster is to the diagonal the shorter its life and therefore the more likely it is to represent random noise rather than an actual cluster.

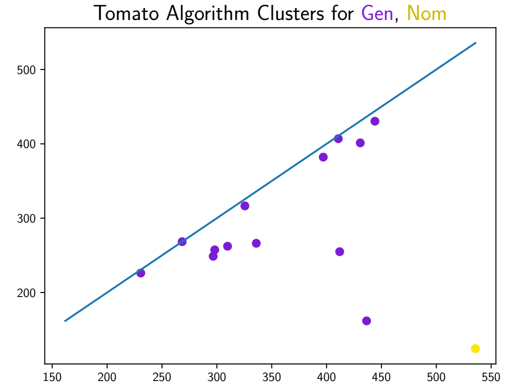


Figure 7: Representation of the ToMATo algorithm for GENITIVE and NOMINATIVE profiles.

In Figure 7 the algorithm proposes multiple clusters, which could not be combined to form better defined clusters. This suggests, as already suggested by Figures 5 and 6 that the possible clusters are not well defined and might overlap with each other. To try and measure the overlap of the clusters, we computed a confusion matrix by the method of the *k-nearest neighbours*.

Target Pred.	Acc	Gen	Loc	Nom
Acc	130	62	51	34
Gen	69	156	16	42
Loc	35	57	29	34
Nom	29	28	9	227

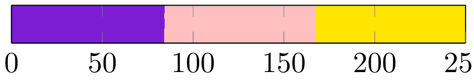


Table 3: Confusion matrix for *k-NN* with $k = 11$ on `Acc, Gen, Loc, Nom`. Rows correspond to the prediction and columns to the expected value.

As we can see in Table 3, while cases that are present in many languages (NOMINATIVE, ACCUSATIVE, GENITIVE) are quite recognisable, it is definitely not obvious, especially when throwing on other less common cases such as locative. In fact, changing the parameter k does not lead to significantly better results. The more common cases are less recognisable with decreasing k, leading to a worse classification, and the less common cases are even more blurred when increasing k, since they are flooded in the total number of samples. Moreover, whatever the parameter, there are always samples from common core cases that are classified as other cases. It appears that the portion of space occupied by each case is neither fully distinct from the others, causing confusion when trying to cluster cases with the same names as well as limiting our ability to distinguish smaller cases from ones that take more space, nor is it well connected, given the fact some samples are always closer to other cases.

7 Adposition Annotation

As discussed in Section 2, some corpora in UD make use of the `Case` feature on adpositions and it is recommended by UD's guidelines.

Given the postulate according to which all natural languages are equally expressive, one could indeed see case marking and the use of adpositions as two means of achieving the same linguistic goals. Two means that are by no mean exclusive since languages that use case tend to have a rather limited inventory and use adpositions to express a broader range of meanings and relations.

Following Kirov et al. (2017), we have applied the methods described above to represent certain adpositions and to give them a syntactically equivalent case representation. This could partially prove the postulate, as well as help justifying the way some corpora annotate adpositions for case.

To do so, we counted the dependency relations leading to the governors of each adposition. This gave us a distribution on syntactic usage of adpositions similar to a profile, and allowed us to compare adpositions to cases.

Table 4 represents the uniform means of the representations of a few French adpositions across all French corpora. As we can see, and could be predicted by French speakers, most adpositions are used in a similar way in French, mainly as LOCATIVES (*dans, par, sur, sous, vers. . .*) or INSTRUMENTALS/COMITATIVE (*avec*). For the other adpositions, we see that there is a non-negligible proportion of usage that leads to `advcl`. This comes from infinitive constructions marking goal (*pour*), intent (*à*), avoidance (*sans*) or gerundive construc-

Adpos	advcl	nmod	nsubj	obj	obl
À	16.7	17.3	0.04	0.38	63.4
DANS	0.46	13.8		0.19	78.7
PAR	0.26	13.7	0.10	0.18	74.6
POUR	29.5	15.9		0.02	41.2
EN	8.13	17.1		0.36	54.1
VERS	0.26	35.7			62.1
AVEC	0.61	32.4			62.6
DE	2.10	68.0	0.14	1.31	14.3
SANS	24.4	21.1		0.78	43.8
SOUS	0.21	22.9	0.02	72.8	
SUR	0.47	36.3		0.10	59.4
SAUF	10.7	22.6			38.1

Table 4: Dependency relation profiles of the governors irrespective of its part-of-speech of a few French adpositions.

tions marking manner (*en*).

This justifies the idea of giving a case to adpositions as a reasonable supposition, and confirms our postulate that adpositions replace some cases in language without cases (French actually has cases on personal pronouns; but not for any of the cases *replaced* by adpositions). We believe that this method could be extended to any other part of speech with adequate semantics and syntactic constructions.

8 Conclusion

In this paper, we have investigated the comparative-descriptive confusion that Haspelmath warned us about using Universal Dependency data. We have compared cases between different languages as is it was a commensurable descriptive category and seen that at least for some closely related languages the alignment stands at least for core cases. We then tried to represent archetypal cases as if case was a comparative concept applied onto each treebank, and saw that core cases mostly align with our expectations. However, this asks for a more principled analysis of the use of the term *nominative* for the default case especially so when the nominative-accusative distinction does not exist or when it does not simply mark a syntactic role but also definiteness for example.

References

Frédéric Chazal, Leonidas Guibas, Steve Oudot, and Primoz Skraba. 2011. Persistence-based clustering in riemannian manifolds. *Journal of the ACM*, 60.

Martin Haspelmath. 2018. *How comparative concepts and descriptive linguistic categories are different*, pages 83–114.

Santiago Herrera, Caio Corro, and Sylvain Kahane. 2024. Sparse logistic regression with high-order features for automatic grammar rule extraction from treebanks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15114–15125, Torino, Italia. ELRA and ICCL.

Christo Kirov, John Sylak-Glassman, Rebecca Knowles, Ryan Cotterell, and Matt Post. 2017. A rich morphological tagger for english: Exploring the cross-linguistic tradeoff between morphology and syntax. pages 112–117.

Vincent Kríž and Barbora Hladká. 2018. Czech legal text treebank 2.0. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Laurens van der Maaten and Geoffrey Hinton. 2008. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Clément Maria, Jean-Daniel Boissonnat, Marc Glisse, and Mariette Yvinec. 2014. The gudhi library: Simplicial complexes and persistent homology. In *Mathematical Software – ICMS 2014*, pages 167–174, Berlin, Heidelberg. Springer Berlin Heidelberg.

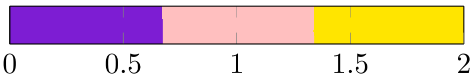
Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Daniel Zeman et al. 2024. Universal dependencies 2.14. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Appendix

List of the dependency relations used for the x-axis in 4:

_; acl; advcl; advmod; amod; appos; aux; case; cc; ccomp; clf; compound; conj; cop; csubj; dep; det; discourse; dislocated; expl; fixed; flat; iobj; list; mark; nmod; nsubj; nummod; obj; obl; orphan; parataxis; punct; reparandum; root; vocative; xcomp



Ru Cs	Acc	Dat	Gen	Ins	Loc	Nom	Par
Acc	0.3	0.45	0.49	0.26	0.51	0.41	1.03
Dat	0.49	0.44	0.55	0.38	0.46	0.51	0.92
Gen	0.5	0.4	0.2	0.32	0.47	0.52	1.05
Ins	0.43	0.39	0.47	0.28	0.42	0.44	0.95
Loc	0.53	0.48	0.54	0.43	0.5	0.55	0.96
Nom	0.49	0.55	0.59	0.38	0.59	0.25	1.11

Table 5: Distances between Czech CLTT and Russian GSD case profiles.

Ru Cs	Acc	Dat	Gen	Ins	Loc	Nom	Par
Acc	0.25	0.63	0.7	0.42	0.78	0.78	1.03
Dat	0.67	0.36	0.64	0.46	0.44	0.83	0.72
Gen	0.92	0.63	0.09	0.64	0.82	1.01	1.17
Ins	0.65	0.42	0.63	0.41	0.55	0.75	0.82
Loc	0.66	0.25	0.49	0.41	0.36	0.84	0.71
Nom	0.81	0.82	0.93	0.68	0.96	0.18	1.16

Table 6: Distances between Czech CLTT and Russian GSD noun case profiles.

	Acc	Dat	Gen	Ins	Loc	Nom
Acc	0	0.32	0.37	0.26	0.35	0.38
Dat	0.32	0	0.39	0.17	0.15	0.51
Gen	0.37	0.39	0	0.32	0.38	0.49
Ins	0.26	0.17	0.32	0	0.26	0.41
Loc	0.35	0.15	0.38	0.26	0	0.56
Nom	0.38	0.51	0.49	0.41	0.56	0

Table 7: Distances between Czech CLTT case profiles.

	Acc	Dat	Gen	Ins	Loc	Nom
Acc	0	0.61	0.76	0.55	0.6	0.71
Dat	0.61	0	0.65	0.18	0.28	0.81
Gen	0.76	0.65	0	0.66	0.5	0.98
Ins	0.55	0.18	0.66	0	0.33	0.74
Loc	0.6	0.28	0.5	0.33	0	0.82
Nom	0.71	0.81	0.98	0.74	0.82	0

Table 8: Distances between Czech CLTT noun case profiles.

	Acc	Dat	Gen	Ins	Loc	Nom	Par
Acc	0	0.44	0.55	0.32	0.46	0.51	0.92
Dat	0.44	0	0.44	0.3	0.27	0.53	0.77
Gen	0.55	0.44	0	0.39	0.51	0.58	1.07
Ins	0.32	0.3	0.39	0	0.39	0.39	0.94
Loc	0.46	0.27	0.51	0.39	0	0.59	0.6
Nom	0.51	0.53	0.58	0.39	0.59	0	1.07
Par	0.92	0.77	1.07	0.94	0.6	1.07	0

Table 9: Distances between Russian GSD case profiles.

	Acc	Dat	Gen	Ins	Loc	Nom	Par
Acc	0	0.65	0.87	0.5	0.72	0.87	0.9
Dat	0.65	0	0.62	0.39	0.34	0.84	0.64
Gen	0.87	0.62	0	0.59	0.82	0.96	1.17
Ins	0.5	0.39	0.59	0	0.61	0.72	0.89
Loc	0.72	0.34	0.82	0.61	0	0.97	0.35
Nom	0.87	0.84	0.96	0.72	0.97	0	1.16
Par	0.9	0.64	1.17	0.89	0.35	1.16	0

Table 10: Distances between Russian GSD noun case profiles.

	Abl	Acc	Dat	Equ	Gen	Ins	Loc	Nom
Acc	0.69	0.62	0.66	0.54	0.74	0.74	0.76	0.43
Dat	0.62	0.83	0.6	0.51	0.78	0.67	0.67	0.52
Gen	0.74	0.86	0.71	0.6	0.81	0.78	0.8	0.57
Ins	0.63	0.81	0.61	0.47	0.76	0.69	0.68	0.49
Loc	0.66	0.85	0.64	0.56	0.81	0.71	0.72	0.57
Nom	0.8	0.81	0.77	0.6	0.73	0.84	0.84	0.43

Table 11: Distances between Czech CLTT and Turkish BOUN case profiles.

	Abl	Acc	Dat	Equ	Gen	Ins	Loc	Nom
Acc	0.86	0.49	0.75	1.14	0.88	0.82	0.91	0.53
Dat	0.57	1.06	0.5	1.16	0.92	0.55	0.6	0.64
Gen	1.03	1.22	0.96	1.3	1.09	1	1.07	0.89
Ins	0.66	1.01	0.57	1.1	0.85	0.64	0.69	0.56
Loc	0.56	1.07	0.49	1.16	0.93	0.53	0.6	0.65
Nom	1.01	0.98	0.91	1.17	0.75	0.99	1.04	0.46

Table 12: Distances between Czech CLTT and Turkish BOUN nouns case profiles.

	Abl	Acc	Dat	Equ	Gen	Ins	Loc	Nom
Abl	0	0.93	0.11	0.4	0.87	0.14	0.12	0.65
Acc	0.93	0	0.87	0.88	0.94	0.96	1.01	0.68
Dat	0.11	0.87	0	0.41	0.88	0.14	0.16	0.63
Equ	0.4	0.88	0.41	0	0.83	0.48	0.44	0.57
Gen	0.87	0.94	0.88	0.83	0	0.94	0.96	0.41
Ins	0.14	0.96	0.14	0.48	0.94	0	0.11	0.71
Loc	0.12	1.01	0.16	0.44	0.96	0.11	0	0.73
Nom	0.65	0.68	0.63	0.57	0.41	0.71	0.73	0

Table 13: Distances between Turkish BOUN case profiles.