Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 109-125 recibido 03-12-2024 revisado 05-02-2025 aceptado 06-02-2025

Morphological Clustering of Neologisms

Agrupación morfológica de neologismos

Jordi Porta-Zamorano, Alejandro Muñoz-Navarro, Carlos Golvano-Díaz, Esther Navío-Castellano

Centro de Estudios de la Real Academia Española

{porta, alejandro.munoz, carlos.golvano, enavio}@rae.es

Abstract: This work addresses the identification of groups of words morphologically related through inflection or derivation. The purpose of this task is to monitor neologisms and consider if they should enter a dictionary depending on evidence gathered. With this aim in mind, a lemmatizer and a derivational analyzer based on deep neural networks have been developed, along with several algorithms from graph theory and complex networks.
Keywords: Neology, lemmatization, derivational analysis, clustering, complex networks.

Resumen: Este trabajo aborda la identificación de grupos de palabras relacionadas morfológicamente mediante flexión o derivación. El propósito de esta tarea es la monitorización de neologismos para poder evaluar si deberían incorporarse a un diccionario según la evidencia recopilada. Con este objetivo, se han desarrollado un lematizador y un analizador derivacional basados en redes neuronales profundas, junto con varios algoritmos de teoría de grafos y redes complejas.
Palabras clave: Neología, lematización, análisis de derivados, agrupamiento, redes complejas.

1 Introduction

Spanish, as a Romance language, has a morphology characterized by its rich inflection system and the use of derivation and compounding to create new words, which are the subject of study of neology. Classifications of neologisms usually include new forms, which might result from affixation (*pantallazo*), compounding (*pagafantas*), different processes of shortening and reduction (*finde* < *fin de semana* 'weekend'), lexicalization of proper names and brands (*vespa*) or, rarely, be *ex nihilo* creations (*suripanta*); loanwords (*feta* 'Greek cheese'); semantic change in preexisting words (*virus* in the sense of 'malicious software'), or changes in category and subcategory, such as gender (*azafato* 'male flight attendant') or verb valency (*topar* 'cap [prices]', transitive verb)—for a wider discussion, see Díaz Hormigo (2011), with a particular focus on the classification proposed by Cabré and Estopà (2004) within the projects promoted by the OBNEO[1]. Finally, innovation might also end up affecting units smaller than a word (*puto-*) or larger (*cinturón verde*).

A lexical family refers to a set of words that share a common segment in their form and have the same initial or remote origin, carrying a basic meaning shared by each unit in the group. This segment or lexical morpheme is the root or radical. For example, in the lexical family formed by *gato*, *gatuno*, *gatera*, *gateo*, *gatear*, the non-decomposable morpheme common to all of them is *gat-*. However, different stages in historical development of the members of a lexical family can erase the formal connection between them, as is the case with *filial*, *filiación*, and *hijo* ('son'), from Latin *filius* or with *episcopal* and *obispo* ('bishop') from Greek through Latin *episcŏpus*. To a speaker without historical knowledge, these cases, where suppletive forms intervene, can be confusing.

[1] https://www.upf.edu/es/web/obneo

ISSN 1135-5948 DOI 10.26342/2025-74-8

This work addresses the problem of clustering unknown single words into morphological-related groups with the aim of providing evidence of lexical families of neologisms to neologists and lexicographers. In this work, the relationships established between words in a group will be restricted to those obtained by inflection and derivation. To achieve this, a machine learning model generates the lemmas of a set of surface words. Next, a second model performs a derivational analysis of the proposed lemmas, connecting derivatives with their bases. Finally, these relationships form a graph where groups can be identified.

The contributions that this article aims to make are, on the one hand, designing and evaluating models and datasets for lemmatization and derivational analysis in Spanish, and, on the other hand, identifying and evaluating techniques for grouping word families, with application to neologism detection.

The paper is structured as follows. First, previous works related to the topics of this study are mentioned. Next, the work done for lemmatization, derivational analysis, and word groups detection is presented. For each task, models, data used, experiments conducted and analysis of the results are detailed, followed by some examples illustrating the application of these techniques. Finally, the paper closes with some conclusions and a discussion of future improvements.

2 Related Work

Most studies on computational morphology addressing tasks such as lemmatization, inflection, reinflection or morphological segmentation have been published in the workshop series of the ACL Special Interest Group on Computational Morphology and Phonology (SIGMORPHON).[2] Before 2019, SIGMORPH shared tasks were dominated by neural sequence-to-sequence models with soft attention, but Wu and Cotterell (2019) showed that enforcing strict monotonicity with hard attention was useful in tasks such as morphological inflection. However, most participating systems in the SIGMORPH 2019 Shared Task 2 (Morphological analysis in context) were based on models such as BERT (Devlin et al., 2019) and perform jointly lemmatization and morphological tagging of sentences.

Today, lemmatization techniques are predominantly based on supervised contextual methods as initially proposed by Chrupała (2006). This approach views the lemmatization process as a classification task in which the surface form and its possible category are classified into a set of classes corresponding to the necessary operations on the form to obtain the lemma. To achieve this, they propose using the shortest edit script (SES) between two strings—the surface form and the lemma—as the class.[3] It consists of a set of instructions (insertions and deletions) that, when applied to a sequence (the surface form), transforms it into another sequence (the lemma). Surprisingly, Toporkov and Agerri (2024b) found that providing lemmatizers models with fine-grained morphological features during training is not more beneficial than providing only the part-of-speech of the word. These results were proven for a relatively wide spectrum of morphologically different languages (Basque, Turkish, Russian, Czech, Spanish and English). In a subsequent article, Toporkov and Agerri (2024a) evaluate three different methods for computing SES and their impact on lemmatization. The best word accuracy results reported for in-domain and out-of-domain settings for Spanish were 98.3% and 97.2% respectively.

With respect to the grouping of morphologically related words, Baroni, Matiasek, and Trost (2002) presented an algorithm that takes an unannotated corpus as its input and returns a ranked list of probable morphologically related pairs as its output. The algorithm tries to discover morphologically related pairs by looking for pairs that are orthographically and semantically similar, where orthographic similarity is measured in terms of minimum edit distance and semantic similarity is measured in terms of mutual information. Experiments with German and English inputs gave encouraging results, considering both precision and the nature of the morphological patterns found within the output set.

In another work, Moon, Erk, and

[2] https://sigmorphon.github.io/

[3] The algorithm for calculating this script was proposed by Myers (1986), and is used by the Un*x utility diff.

Baldridge (2009) presented a model to generate *conflation sets*, defined as sets of word types that are related through either inflectional or derivational morphology. Their approach is a four-stage process: (1) candidate stems and affixes generation using a trie data structure; (2) candidate affixes filtering using statistical significance for pairs of affixes based on their co-occurence counts with shared stems; (3) significant affix pairs clustering into affix groups; and (4) conflation sets generation based on affix clusters. They applied this method to English and Uspanteko, an endangered Mayan language. Their model compared favorably to two well-known systems for unsupervised morphology acquisition: Morfessor (Creutz and Lagus, 2007) and Linguistica (Goldsmith, 2001). The work contains also a summary of the diverse body of existing work at that time on morphological segmentation and clustering.

3 Lemmatization Task

Spanish words, especially verbs, nouns, and adjectives, are extensively inflected to convey grammatical information and mark grammatical relationships. Verbs are inflected for tense, aspect, mood, person and number. Spanish has clitic pronouns (e.g., *me*, *te*, *se*, *lo*, *la*...), which correspond mainly to direct or indirect objects, though they also appear in pronominal verbs and in the use of impersonal *se*. These clitics precede inflected verbs (*lo veo* 'I see him'), but are attached to imperatives and non-finite verbs (*verlo* [see-INF-3rd.masc.sing.acc] 'to see him').

3.1 Data

We have used as base lexicon the full-form lexicon used at the Royal Spanish Academy to annotate corpora like CORPES XXI[4]. This lexicon relates word-forms and morpho-syntactic descriptions to their corresponding lemmas in the *Diccionario de la lengua española* (RAE and ASALE, 2014) and in the *Diccionario de americanismos* (ASALE, 2010). This base lexicon has been augmented with other words found in the CORPES XXI corresponding to verbal forms with enclitics (verb forms with pronouns attached to the end).

Lemmatization has been adapted for tasks such as neologism detection. We have removed the *se* pronoun attached to the end of the lemma of pronominal verbs (*enorgullecerse* → *enorgullecer*) and the inflective paradigm of nouns and adjectives has been extended to capture intended gender-neutral marks (*-e*, *-x*, *@*). Other ways of explicitly referring to both genders, such as *bienvenido/a* or *bienvenido(a)*, have also been considered, as well as number variations like in *bienvenido(s)*.

The complete dataset contains 976,413 examples, comprising 887,066 unique surface forms and 118,163 distinct lemma and part-of-speech combinations.

3.2 Models and Experiments

Yoyodyne[5] provides neural models for small-vocabulary sequence-to-sequence generation with and without feature conditioning. For lemmatization, we have used the part-of-speech tag of a word as a conditioning signal for inferring its lemma. Yoyodyne provides classic LSTM and transformer models, including hard attention and pointer generation models.

The transformer architecture (Vaswani et al., 2017) achieved state-of-the-art performance on a wide range of word-level sequence-to-sequence tasks, but recurrent models with attention historically dominated character-level transduction tasks (Cotterell et al., 2018). However, Wu, Cotterell, and Hulden (2021) found that, in tasks such as morphological inflection or historical normalization, transformers can outperform recurrent models when they are trained with batch sizes greater than 128, and that this late result went unnoticed because typical datasets for character-level transductions were rather small (10k examples) compared to datasets for word-level tasks. For inflection, they did not find any phenomenon (such as word-internal sound changes, vowel harmony, circumfixation, ablaut or umlaut) that gives a consistent advantage to transformers over recurrent models. Taking these previous results into account, we have selected the transformer and LSTM with attention architectures to conduct the experiments on lemmatization and derivational analysis.

[4]Corpus del Español del Siglo XXI: `https://www.rae.es/corpes`

[5]`https://github.com/CUNY-CL/yoyodyne`

	Accuracy (%)								
	Transformer		Attentive LSTM						
PoS	**G**	**H**	**G**	**H**	**@1**	**@2**	**@3**	**@4**	**Total**
Verbs	97.340	98.308	97.224	98.244	97.232	99.127	99.450	99.580	52,306
Adjs.	97.303	99.071	96.974	98.931	97.017	99.284	99.676	99.791	56,849
Nouns	89.177	96.505	88.940	96.725	88.988	98.004	99.209	99.564	202,959
Total	95.966	98.145	95.790	98.115	95.811	98.968	99.451	99.616	312,114

Table 1: Word lemmatization accuracy for each part of speech (PoS) and model architecture. Column G represents the results with the most frequent lemma produced by the model with a greedy search decoder. Column H shows the results with a heuristically engineered decoder. In contrast, Columns @k present the accuracy considering the one to k most probable candidates delivered by a beam search decoder with a beam of four. Beam decoding was not available for the transformer. The column Total contains the number of examples for each word class in the test set and the row Total contains the micro-averaged accuracy.

Having sufficient examples of each part of speech to split into training (60%), validation (20%) and test sets (20%), we opted to sample the lemma set, ensuring that the inflected forms of a lemma do not appear in more than one partition and that no lemma contamination could affect the significance of accuracy metrics.

For the transformer and the attentive LSTM, Table 1 shows accuracy reached for each part of speech considered using a greedy search decoder (G), a heuristically engineered decoder (H) generating different hypotheses based on errors observed, and a decoder performing a beam search with a width of four for the LSTM. The heuristic decoder generates different alternatives based on the word's ending and its word class. When it is a noun or an adjective, it changes the final vowel (*-a*, *-o*, *-e*) or replaces it with ∅. In the case of verbs, it generates alternatives with all three conjugations (*-ar*, *-er*, *-ir*).

3.3 Analysis of Errors

Regarding errors in nominal inflection, a significant number of errors are observed in the ending vowels of the predicted lemmas, likely due to overgeneralizations of the endings *-a* and *-o* as markers of feminine and masculine gender. The model fails to produce lemmas of inherently feminine nouns that end in *-a* (*contraseña*, *cremallera*), as well as of nouns that end in *-e* (*laringe*, *epígrafe*) or a consonant (*luz*, *carácter*). In the latter group, challenges are particularly evident in the spelling of /*θ*/ before vowels *e* and *i*. When the feminine gender marker exhibits morphological increment, as is the case with *-esa*, the model tends to take only the final *-a* as gender marker and thus consider the increment part of the lemma. Hence, given *abadesa*, it would predict *abadés*, instead of *abad*, along the pattern of *marquesa*/*marqués*. Some other errors are also due to homography. For example, given *cuerdas*, both *cuerda* ('rope') and *cuerdo* ('sane') would be correct predictions, but for each given form only one prediction can be made. An overgeneralization of *-s* and *-es* as plural markers can also be attested. Both the transformer and the LSTM fail to predict lemmas ending in these segments (for example, given *atlas*, **atla* is suggested). Therefore, the same happens systematically with *pluralia tantum* nouns (*fauces*, *fastos*), lexicalized words (*agonías*, *botones*) and verb-noun compounds (*cuentacuentos*, *lanzamisiles*).

For verbal lemmatization, main challenges observed include the difficulty of assigning a conjugation ending (*-ar*, *-er*, *-ir*) to the lemma when the theme vowel in verbal forms used as input is covert. For example, given *carduce*, the lemma predicted is *carducir*, as if it were third person singular indicative present, along the model of *conducir*, instead of *carduzar*, as if it were first or third person singular subjunctive present, like *desmenuzar*. For a verb as rare as *carduzar*, hesitation between the two conjugations is plausible among native speakers of Spanish, since this plurality of analyses is frequent in the language (*sume* can be indicative

present of *sumar* or subjunctive present of *sumir*). Irregular or highly irregular verbs, such as *haber* or *ir*, as well as vowel-final stem verbs, which involve unpredictable accent or final vowel changes (*anunciar*, *averiguar*, *adecuar*), led to a number of errors. Finally, regular conjugated present forms also produced wrong predictions, particularly when the final segments of a stem coincided with an infinitive ending. Thus, given *admira* and *tolera*, **admir* and **toler* were predicted, instead of *admirar* and *tolerar*.

These and the preceding examples where the system fails are quite understandable, as there are no clues in the word-form to predict its ending. Ending vowels in Spanish nouns have their origins in Latin; their current form is the result of a long process of phonetic evolution from Vulgar Latin to the Romance languages and eventually to modern Spanish. This process included the elimination of distinctions between long and short vowels, which implied a readjustment of the vowel system. We observed that the beam search decoder produces most of the hypotheses generated by the heuristic decoder, as in the following example, where the higher the score, the better:

```
sociópatas + A
→ sociópata   -0.155339494347572
→ *sociópato  -2.199321269989013
→ *sociópatas -3.698281526565551
→ *sociópate  -5.522466659545898
```

The quantitative analysis of results shown in Table 1 reveals that the transformer outperforms the LSTM model in each part of speech using a greedy decoder. However, the LSTM outperforms the transformer using beam search. For the LSTM, it can also be seen that decoding with heuristics (producing variations on the lemma similar to those observed in error analysis) improves results compared to greedy decoding, especially for nouns, but does not perform as well as decoding the LSTM with a beam of two. However, if we consider decoding time, using beam search is costly because inference—at least with the current implementation—must be done example by example, without the possibility of leveraging the computation on the GPU.

4 Derivational Analysis Task

Spanish also uses derivational morphemes (mainly prefixes, suffixes and circumfixes) to create new words or adding particular connotations to existing words, as is the case with appreciative suffixation. Taking into account previous works on morpholexical relations (Santana et al., 2003) and present-day Spanish grammars (Bosque and Demonte, 1999; RAE and ASALE, 2009-2011), we have adopted the *Diccionario de afijos del español contemporáneo* (DAEC), by A. Fábregas (2023), as a reference for its value as a comprehensive and systematic inventory of affixes. The dictionary describes the main properties of contemporary Spanish affixes, including their grammatical behavior, base classes, meanings, phonological behavior, and relation to other affixes.

4.1 Data

Affixes from the DAEC have been double-checked against the affix frequency list in Cantos Gómez and Almela Pérez (2009) and Moliner (2013), among other works, to ensure the inclusion of the most productive affixes in the training set.

Along with the affix, the word class of its base is encoded: *N* for nouns and adjectives; *V* for verbs. We have compiled 237 affixes with at least 10 examples each: 96 prefixes (*des-V*, *des-N*, *anti-N*...), 128 suffixes (*V-dor*, *V-ción*, *N-ero*...) and 13 circumfixes (*en-N-ar*, *a-N-ar*, *a-N-ear*...). The total number of triplets consisting of derivative, affix, and base is 48,948.

4.2 Models and Experiments

We have used the same models architectures as for lemmatization to approach derivational analysis. In this case, the model is trained using triplets that represent the input word, the affix and the output (e.g., ⟨*comedor*, *V-dor*, *comer*⟩), where the affix is used as conditioning.

Stratified sampling is applied to ensure adequate representation of all affixes across training, validation and test partitions. However, due to the unbalanced number of examples for each affix (not all are equally productive), a bias could affect the model's accuracy measurement. To mitigate this, we repeated the experiment ten times to obtain its average accuracy and standard deviation.

As for lemmatization, we have performed experiments with different decoders for the two models: the greedy, the heuristic and the beam search decoder for the LSTM.

Like in the lemmatization task, we observed that variation in the candidates suggested by the model are linguistically plausible, like the following:

```
anudar + a-N-ar
→ nudo -0.382106274366378
→ *nuda -1.466021060943603
→ *nud  -3.548364162445068
→ *núdo -4.309681415557861
```

Model	Decoder	Accuracy (%)
Transfomer	G	0.8033 ± 0.0051
	H	0.8964 ± 0.0041
Attentive LSTM	G	0.8349 ± 0.0113
	H	0.9143 ± 0.0059
	@1	0.8309 ± 0.0143
	@2	0.9207 ± 0.0076
	@3	0.9510 ± 0.0056
	@4	0.9573 ± 0.0140

Table 2: Derivational analysis average accuracy and standard deviation of 10 runs using different decoders: a greedy search decoder (G), a heuristic decoder (H), and a beam search decoder with a beam width of four considering the one to k most probable candidates (@k).

4.3 Analysis of Errors

The accuracy results from applying the same decoders used for lemmatization are shown in Table 2. As for lemmatization, the improvement in precision can also be seen when applying the same simple heuristics or the top-k best results from a beam search. Apart from the difficulty to predict the ending vowel in nouns observed in the previous task, partially due to an overgeneralization of *-o/-a* alternations (for example, given *prolífico* ('prolific'), **prolo* is predicted, instead of *prole*, some other difficulties arise because of the nature of the task. For example, if the derivative results in monophthongization (*vejez* 'old age'), the lemma predicted preserves it (**vejo*), when the actual base has a diphthong (*viejo* 'old person'). In some cases, both monophthongized and diphthongized bases are acceptable; hence, predicted *sinvergüenzón* is as valid as expected *sinvergonzón*, but, again, only one prediction can be made. This also applies to nouns with two orthographically valid variants, such as *frijol/fríjol*, *folclor/folclore* or verbal bases vacillating between the second and the third conjugation *emerger/emergir*.

The use of diacritics is also challenging, particularly when the orthographical context justifying its use in the derivative disappears in the base. Thus, diaeresis is incorrectly preserved when, given *yegüero*, the model predicts *yegüa*, not *yegua*. The same happens with accent marks. The derivative *antigás* must carry an accent mark, but its base *gas*, a monosyllabic word, should not; however, the model predicts **gás*.

Again, orthographic conventions affecting the representation of /*θ*/ before vowels *e* and *i* lead to wrong predictions (*pancismo* > **panco* instead of *panza*). Haplology and segment reduction attested in derivation also proved difficult, since those deleted segments should be restored in the process of retrieving the base (*supersticioso* > **supersticio* instead of *superstición*; **supersticionoso* would be the resulting form if haplology did not occur).

Finally, deverbal adjectives ending in *-nte* can also lead to wrong predictions, since verbs of the second conjugation can produce either adjectives ending in *-ente* (*sorprendente* < *sorprender*) or *-iente* (*contendiente* < *contender*); *-iente* is associated to adjectives formed from verbs of the third conjugation (*conveniente* < *convenir*).

In general, it can be stated that the model applies a conservative strategy when identifying the most probable base of a derivative. That said, the expected base is often found among the top most probable bases proposed by the model.

5 Clustering Task

The first step in finding morphologically-related groups within a graph involves identifying its connected components, i.e., the sets or subgraphs with vertices connected to each other. These subgraphs are too coarse, as they often group multiple derivational families together because of some coincidental affixation relationship between them. In graph theory, bridges or cut-edges are edges whose removal increases the number of connected

components. However, the spurious connections generated by derivational analysis created so much noise that the bridges-based approach did not work as expected. More robust solutions involve using community detection algorithms designed for complex networks, such as the algorithms described in the next section.

5.1 Community Detection Algorithms

We applied a community detection algorithm to each connected component. Specifically, we used the edge betweenness-based detection algorithm, which aims to find the structure of communities or modules in a network by identifying the edges that are part of the most shortest paths between any pair of nodes in the network. In particular, the algorithm of Girvan and Newman (2002) progressively removes edges with the highest betweenness value until maximum modularity is reached in the communities detected. This modularity value is based on the density of connections within nodes of a community versus those of other communities. In addition, we applied the community detection algorithm presented in Reichardt and Bornholdt (2006). This algorithm interprets the network's community structure as the spin configuration minimizing the energy of a spin glass model (a Hamiltonian function), where the spin states correspond to community assignments. The algorithm uses a tuning parameter γ that controls the relative importance of internal connections within communities compared to connections between different communities. Higher γ values increase the emphasis on separating communities, resulting in smaller and more numerous communities. Lower γ values make the algorithm favor grouping more nodes together, resulting in fewer but larger communities.

5.2 Data and Experiments

For evaluating the performance of community detection algorithms, we used the list of the 10,000 most frequent lemmas of CORPES XXI[6], from which we selected nouns (4,726), adjectives (1,853) and verbs (1,656). The morphological-lexical families were manually identified for the 27 largest connected components, accounting for 449 words.

The manual identification of families is not without theoretical issues, since different criteria such as etymology or synchronic speaker perception can be used to split components into families. We finally prioritised the creation of families based on etymological criteria. During annotation, some families were separated into different components while families corresponding to homographs such as *canto* 'edge' and *canto* 'chant' converged into a single component.

To evaluate the algorithms, we used Homogeneity, Completeness, and V-Measure as defined by Rosenberg and Hirschberg (2007). Homogeneity checks if each community contains only members of a single group, completeness checks if all members of a given group are in the same community, and V-measure is their harmonic mean. Table 3 contains the results for several algorithms, including the metrics applied to the connected components with no further refinement. However, while the time complexity of the connected components algorithm is linear in the number of vertices and edges, the Spinglass algorithm uses simulated annealing to approximate a solution. In general, simulated annealing does not have a guaranteed polynomial-time complexity for all problems, since time complexity may depend on the complexity of the search space. This means that, for some problems, the time to find a good solution may grow exponentially with the problem size.

Algorithm	H	C	V
Conn. components	0.914	1.000	0.955
Edge-betweenness	0.971	0.752	0.847
Spinglass (γ=1)	0.976	0.735	0.838
Spinglass (γ=0.001)	0.924	0.996	0.958

Table 3: Metrics for different community detection algorithms: Homogeneity (H), Completeness (V) and V-Measure (V). Edge-betweenness and Spinglass are applied on the components of the network.

5.3 Analysis of Results

As can be seen in Table 3, the etymological criterion used for identifying families manually tends to favour bigger groups. This statement is supported by the fact that the homogeneity of connected components is

[6] https://www.rae.es/corpes/assets/rae/files/corpes/10000_lemas.html

high (H=0.914). By contrast, algorithms for community detection improve homogeneity at the expense of completeness, i.e., they separate members of the same group into different communities. Looking at the V-measure as a global metric and conducting a systematic search on the hyperparameter γ at different scales, it seems that using the best γ value (0.001) for Spinglass produces results similar to connected components without further refinement.

Out of the 27 groups, 16 include members of a single family, sharing one root (*viv-* in *vivo, vivir, viviente, revivir...*) or suppletive roots (*pos-* in the derivatives of *poner*, such as *componer, exponer, posición...*). In other cases, derivatives etymologically related have split into two main groups. This is the case with words ultimately related to *tendere* ('to stretch', 'to tend'), reflecting /d/∼/s/ consonant alternation in the root (*tendere*, passive participle *tensus*). Hence, one component includes *pretender, entender, extender...*, whereas another one includes *tenso, tensión, pretensión....* This division is consistent with a synchronic perspective, since *tensión* was not derived within Spanish. The exceptional alternation /d/∼/θ/ in *atención* accounts for its inclusion into a component with derivatives from *tener* (*detener, detención, retener, retención...*) and *contar* (*contador, contable, cuento...*), instead of appearing in the second component around *tender*. This might have been induced by *contención*, which matches both the beginning of some derivatives of *contar* and the end of certain derivatives of *tener*. Finally, the impossibility to discriminate between semantically different but formally identical roots also results in the inclusion of different families into one component, as is the case with the derivatives of *puerta* ('door'), *puerto* ('port') and *portar* ('to carry') or of the derivatives of the above mentioned *canto* and *canto*, which have also been grouped with unrelated *cantidad* ('quantity').

6 Results on Unknown Words

To get an idea of the performance of the different modules in running texts, in which neologisms may appear, we have applied this methodology to words labelled as unknown in CORPES XXI, i.e. words that have no entry in the corpus analysis lexicon. Once lemmatization and derivational analyses have been performed, the next step is to build the network and decompose it into connected components. The size of connected components seems to follow a power-law distribution, with a giant component that groups 728 lemmas containing several groups of words, accidentally connected because they share certain letter sequences, though they are morphologically unrelated.

Table 4 contains some examples of the connected components found by the algorithm. Component (1) corresponds to a group for the non-standard *güevo* (*huevo*, 'egg'), including spelling errors. Component (2) shows a group of words related to *bonche* (from English *bunch*), included in several Americanisms dictionaries with different meanings ('group of people', 'fuss, riot', 'noisy celebration', among others). Component (3) captures derivatives of *pegote* not included in dictionaries. Component (4) contains a group of lemmas and inflected forms of erroneous **inflingir*, a common mistake confusing *infringir* ('to infringe') and *infligir* ('to inflict'). Finally, component (5) groups some derivatives (not all of them well-formed) of the brand name *Google*. However, the algorithm includes some other members in separate singleton clusters ($googlazo_N$, $googlelazo_N$, $googlero_N$, $googlizado_A$); according to Spanish orthographic conventions, *oo* should be replaced by *u*.

7 Conclusions and Future Work

Word formation is concerned with analysing and understanding the mechanisms by which words are created and evolve. In this work, we focused on the morphological mechanisms of inflection and derivation; compounding, including neoclassical themes, will be addressed in the future, as well as other mechanisms, such as borrowing, initialisms, clipping or blending, whose modelling poses significant challenges.

Despite the apparent success in lemmatization and derivational analysis, the analysis of errors trying to identify the strengths and weaknesses of our approach has provided, like in Gorman et al. (2019), insights for implementing heuristics and future improvements. On the one hand, this analysis allows us to establish the upper limits

Id.	Connected Component
(1)	*agüevada*$_N$ *agüevado*$_N$ *agüevado*$_A$ *agüevar*$_V$ *agüevonado*$_A$ *agüevonear*$_V$ *güevada*$_N$ *güeva*$_N$ *güevo*$_N$ *güevá*$_N$ *güevar*$_V$ *güevazo*$_N$ *güeveo*$_A$ *güeveo*$_N$ *güevera*$_N$ *güevero*$_N$ *güeviar*$_V$ *güevito*$_N$ *güevón*$_A$ *güevonada*$_N$ *güevón*$_N$ *güevoncito*$_N$ *güevonear*$_V$ *güevon*$_N$ *güevudo*$_A$
(2)	*bonche*$_N$ *bonchar*$_V$ *bonchear*$_V$ *boncheo*$_N$ *bonchón*$_A$ *bonchón*$_N$
(3)	*apegotonado*$_A$ *empegotar*$_V$ *empegotado*$_A$ *pegotar*$_V$ *pegotón*$_N$
(4)	*autoinflingir*$_V$ (*autoinflingía*) *autoinflingido*$_A$ (*autoinflingida*, *autoinflingido*) *inflingir*$_V$ (*Inflingen*, *inflingeron*, *inflinge inflingían*, *inflingía*, *inflingida*, *inflingido*, *inflingido*, *inflingiendo*, *inflingieron*, *inflingió*, *inflingirle*, *inflingir*), *inflingido*$_A$ (*inflingido*).
(5)	*googleable*$_A$ *googlear*$_V$

Table 4: Examples of connected components of the network of the CORPES XXI unknown words.

of lemmatization and derivational analysis tasks on general vocabulary. On the other hand, having obtained good results in the lemmatization task, it would be possible to evaluate other lemmatizers, such as UDPipe (Straka, 2018), IXAPipe (Agerri, Bermudez, and Rigau, 2014) or Toporkov and Agerri (2024a), under the same experimental conditions.

It is expected that some of the modules and the methodology presented in this work will be integrated into the Observatory of Words currently under development within the LEIA project.

Acknowledgments

We would like to give special thanks to Antonio Fábregas for sharing his dictionary of affixes in part of this work. Also Kyle Gorman and Adam Wiemerslage deserve our thanks for helping us to make beam search available in Yoyodyne. We also extend our gratitude to the anonymous reviewers for their comments and suggestions, which have contributed to the improvement of this work. This publication results from work undertaken within the framework of the LEIA Project, promoted by the Spanish Ministry of Economic Affairs and Digital Transformation and by the Recovery, Transformation, and Resilience Plan, funded by the European Union - NextGenerationEU.

References

Agerri, R., J. Bermudez, and G. Rigau. 2014. IXA pipeline: Efficient and ready to use multilingual NLP tools. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3823–3828, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

ASALE. 2010. *Diccionario de americanismos*. Santillana.

Baroni, M., J. Matiasek, and H. Trost. 2002. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 48–57. Association for Computational Linguistics, July.

Bosque, I. and V. Demonte, editors. 1999. *Gramática descriptiva de la lengua española*. Espasa.

Cabré, M. and R. Estopà. 2004. Metodología del trabajo en neología: criterios, materiales y procesos. Technical report, Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada.

Cantos Gómez, P. and R. Almela Pérez. 2009. Estudio cuantitativo de los afijos en español. *The Bulletin of Hispanic Studies*, (4):453–468.

Chrupała, G. 2006. Simple data-driven context-sensitive lemmatization. *Procesamiento de lenguaje natural*, (37):121–127.

Cotterell, R., C. Kirov, J. Sylak-Glassman, G. Walther, E. Vylomova, A. D. McCarthy, K. Kann, S. J. Mielke, G. Nicolai, M. Silfverberg, D. Yarowsky, J. Eisner, and M. Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In M. Hulden and R. Cotterell, editors, *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels, October. Association for Computational Linguistics.

Creutz, M. and K. Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, 4(1), February.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Díaz Hormigo, M. T. 2011. Word formation processes and proposals for the classification of formal neologisms. In J. L. Cifuentes and S. Rodríguez, editors, *Spanish word formation and lexical creation*, pages 347–368. John Benjamins.

Fábregas, A. 2023. *Diccionario de afijos del español contemporáneo*. Routledge.

Girvan, M. and M. E. J. Newman. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Goldsmith, J. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198, 06.

Gorman, K., A. D. McCarthy, R. Cotterell, E. Vylomova, M. Silfverberg, and M. Markowska. 2019. Weird inflects but OK: Making sense of morphological generation errors. In M. Bansal and A. Villavicencio, editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 140–151, Hong Kong, China, November. Association for Computational Linguistics.

Moliner, M. 2013. *Neologismos del español actual*. Gredos.

Moon, T., K. Erk, and J. Baldridge. 2009. Unsupervised morphological segmentation and clustering with document boundaries. In P. Koehn and R. Mihalcea, editors, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 668–677, Singapore, August. Association for Computational Linguistics.

Myers, E. W. 1986. An O(ND) Difference Algorithm and Its Variations. *Algorithmica*, 1(2):251–266.

RAE and ASALE. 2009-2011. *Nueva gramática de la lengua española*. Espasa.

RAE and ASALE. 2014. *Diccionario de la lengua española*. Espasa.

Reichardt, J. and S. Bornholdt. 2006. Statistical mechanics of community detection. *Phys. Rev. E*, 74:016110, Jul.

Rosenberg, A. and J. Hirschberg. 2007. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In J. Eisner, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic, June. Association for Computational Linguistics.

Santana, O., F. J. Carreras, J. R. Pérez, and G. Rodríguez. 2003. Relaciones morfoléxicas sufijales del español. *Procesamiento del Lenguaje Natural*, 30.

Straka, M. 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.

Toporkov, O. and R. Agerri. 2024a. Evaluating shortest edit script methods for contextual lemmatization. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti,

and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6451–6463, Torino, Italia, May. ELRA and ICCL.

Toporkov, O. and R. Agerri. 2024b. On the Role of Morphological Information for Contextual Lemmatization. *Computational Linguistics*, 50(1):157–191, 03.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wu, S. and R. Cotterell. 2019. Exact hard monotonic attention for character-level transduction. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537, Florence, Italy, July. Association for Computational Linguistics.

Wu, S., R. Cotterell, and M. Hulden. 2021. Applying the transformer to character-level transduction. In P. Merlo, J. Tiedemann, and R. Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online, April. Association for Computational Linguistics.

A Annex 1: Derivational Analysis Results

The following table shows the results of derivational analysis by affixes averaged from ten runs using the attentive LSTM model. In the table, Affix is the column that contains the label in which, in addition to the affix, the part of speech of the base is represented and serves as a signal for the derivational analyzer. The columns Train, Devel, and Test contain the number of examples in the dataset partitions, which add up to a total expressed in the Total column. The rest of the columns contain the accuracies of different decoding strategies: greedy (G), heuristic (H) and considering the k most probable candidate in a beam of five (@1 to @5).

Affix	Total	Train	Devel	Test	G	H	@1	@2	@3	@4	@5
V-dor	2795	2236	112	447	0.98±0.00	0.98±0.00	0.98±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00
V-ado	2443	1955	98	390	0.99±0.00	0.99±0.00	0.99±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
V-ción	2443	1955	98	390	0.94±0.01	0.94±0.01	0.94±0.00	0.96±0.00	0.96±0.00	0.97±0.00	0.97±0.00
V-miento	1754	1404	70	280	0.98±0.00	0.99±0.00	0.98±0.01	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00
N-ero	1428	1143	57	228	0.63±0.03	0.88±0.01	0.64±0.06	0.87±0.03	0.94±0.01	0.96±0.01	0.96±0.01
N-ía	1420	1136	57	227	0.86±0.03	0.89±0.02	0.85±0.02	0.91±0.02	0.94±0.01	0.96±0.01	0.96±0.01
N-idad	1351	1081	54	216	0.92±0.01	0.93±0.01	0.93±0.00	0.97±0.01	0.98±0.00	0.98±0.00	0.98±0.00
N-al	1237	990	50	197	0.65±0.02	0.85±0.01	0.67±0.02	0.85±0.01	0.91±0.01	0.94±0.01	0.94±0.01
N-ico	1167	934	47	186	0.64±0.05	0.71±0.04	0.64±0.03	0.77±0.02	0.84±0.02	0.88±0.02	0.88±0.02
N-ear	1164	932	47	185	0.61±0.07	0.86±0.03	0.60±0.05	0.84±0.02	0.93±0.01	0.95±0.00	0.95±0.00
N-ísimo	1091	873	44	174	0.90±0.03	0.91±0.03	0.90±0.00	0.96±0.01	0.98±0.00	0.98±0.00	0.98±0.00
V-nte	1079	864	43	172	0.98±0.01	0.99±0.00	0.98±0.01	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00
N-oso	1039	832	42	165	0.58±0.05	0.86±0.02	0.54±0.05	0.82±0.02	0.89±0.01	0.92±0.00	0.92±0.00
N-ismo1	984	788	40	156	0.68±0.04	0.80±0.03	0.70±0.03	0.81±0.03	0.86±0.02	0.87±0.02	0.87±0.02
V-ble	973	779	39	155	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00
en-N-ar	949	760	38	151	0.59±0.04	0.84±0.02	0.58±0.04	0.84±0.04	0.92±0.01	0.94±0.01	0.94±0.01
des-V	933	747	38	148	0.97±0.00	0.97±0.00	0.98±0.00	0.99±0.01	0.99±0.01	0.99±0.00	0.99±0.00
N-era	880	704	36	140	0.65±0.03	0.84±0.02	0.62±0.03	0.83±0.03	0.92±0.02	0.95±0.01	0.95±0.01
a-N-ar	860	688	35	137	0.59±0.04	0.87±0.04	0.59±0.05	0.84±0.01	0.91±0.01	0.93±0.01	0.93±0.01
des-N	854	684	34	136	0.95±0.01	0.95±0.01	0.95±0.02	0.98±0.01	0.98±0.01	0.98±0.01	0.98±0.01
N-ista1	813	651	33	129	0.71±0.03	0.87±0.03	0.72±0.05	0.86±0.03	0.92±0.02	0.93±0.03	0.93±0.03
N-ería	704	564	28	112	0.67±0.02	0.88±0.02	0.68±0.07	0.88±0.03	0.95±0.01	0.96±0.01	0.96±0.01
N-ón	686	549	28	109	0.61±0.04	0.88±0.03	0.59±0.04	0.85±0.03	0.92±0.02	0.94±0.01	0.94±0.01
in-N	647	518	26	103	0.99±0.00	0.99±0.00	0.99±0.01	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00
V-dura	629	504	25	100	0.98±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00
N-ada	574	460	23	91	0.69±0.05	0.86±0.03	0.69±0.07	0.85±0.04	0.93±0.02	0.94±0.02	0.94±0.02
re-N	554	444	22	88	0.98±0.01	0.98±0.01	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00
V-tivo	534	428	22	84	0.84±0.02	0.85±0.02	0.82±0.03	0.87±0.03	0.89±0.02	0.90±0.02	0.90±0.02
V-ido1	476	381	19	76	0.90±0.03	0.98±0.01	0.91±0.02	0.99±0.01	0.99±0.00	0.99±0.00	0.99±0.00
re-V	458	367	19	72	0.98±0.02	0.99±0.01	0.99±0.00	0.99±0.01	0.99±0.00	0.99±0.00	0.99±0.00
N-ura	429	344	17	68	0.87±0.03	0.92±0.02	0.88±0.03	0.94±0.02	0.96±0.02	0.96±0.01	0.96±0.01
V-se	427	342	17	68	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
N-azo	426	341	17	68	0.67±0.05	0.88±0.04	0.67±0.05	0.89±0.03	0.94±0.02	0.96±0.01	0.96±0.01

Table 5 – continued on next page

Affix	Total	Train.	Devel.	Test	G	H	@1	@2	@3	@4	@5
anti-N	420	336	17	67	0.97±0.02	0.97±0.02	0.94±0.08	0.96±0.06	0.96±0.06	0.96±0.07	0.96±0.07
N-izar	418	335	17	66	0.68±0.04	0.75±0.05	0.71±0.04	0.82±0.05	0.85±0.05	0.87±0.05	0.87±0.05
N-ar	407	326	17	64	0.58±0.03	0.86±0.04	0.56±0.03	0.85±0.04	0.91±0.02	0.95±0.02	0.95±0.02
N-illo	403	323	16	64	0.82±0.03	0.91±0.02	0.78±0.05	0.91±0.03	0.95±0.02	0.96±0.02	0.96±0.02
V-dero	340	272	14	54	1.00±0.00	1.00±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00
pre-N	288	231	12	45	0.99±0.01	0.99±0.01	0.98±0.02	0.98±0.02	0.99±0.01	0.99±0.01	0.99±0.01
V-eo	284	228	12	44	0.93±0.02	0.93±0.02	0.91±0.05	0.96±0.05	0.97±0.05	0.97±0.05	0.97±0.05
sub-N	284	228	12	44	0.99±0.01	0.99±0.01	0.98±0.02	0.98±0.02	0.98±0.02	0.98±0.02	0.98±0.02
N-illa	271	217	11	43	0.82±0.07	0.90±0.03	0.87±0.06	0.94±0.03	0.96±0.02	0.97±0.02	0.97±0.02
auto-N	260	208	11	41	0.98±0.01	0.98±0.01	0.96±0.04	0.97±0.03	0.97±0.02	0.97±0.02	0.97±0.02
V-ncia	258	207	11	40	0.79±0.07	0.82±0.06	0.78±0.08	0.87±0.05	0.89±0.05	0.92±0.04	0.92±0.04
V-ón	243	195	10	38	0.94±0.03	0.95±0.03	0.92±0.03	0.96±0.03	0.96±0.03	0.96±0.03	0.96±0.03
N-ismo2	243	195	10	38	0.99±0.01	0.99±0.01	0.98±0.01	0.99±0.01	0.99±0.01	0.99±0.01	0.99±0.01
N-ista2	243	195	10	38	0.98±0.02	0.98±0.02	0.98±0.02	0.99±0.01	0.99±0.01	0.99±0.01	0.99±0.01
V-torio	241	193	10	38	0.96±0.02	0.96±0.03	0.96±0.01	0.99±0.01	0.99±0.01	0.99±0.01	0.99±0.01
sobre-N	222	178	9	35	0.97±0.03	0.97±0.03	0.96±0.02	0.98±0.02	0.98±0.02	0.98±0.02	0.98±0.02
contra-N	215	172	9	34	0.93±0.04	0.93±0.04	0.94±0.02	0.95±0.02	0.95±0.02	0.95±0.02	0.95±0.02
N-eta	211	169	9	33	0.71±0.06	0.87±0.05	0.64±0.07	0.87±0.05	0.93±0.03	0.95±0.03	0.95±0.03
en-N-ado	208	167	9	32	0.64±0.10	0.86±0.06	0.60±0.11	0.82±0.07	0.87±0.08	0.88±0.06	0.88±0.06
micro-N	200	160	8	32	0.99±0.01	0.99±0.01	0.97±0.02	0.98±0.02	0.98±0.02	0.98±0.02	0.98±0.02
N-ito	199	160	8	31	0.89±0.05	0.93±0.03	0.89±0.05	0.96±0.02	0.98±0.02	0.98±0.02	0.98±0.02
V-dera	198	159	8	31	0.99±0.01	0.99±0.01	0.99±0.01	0.99±0.01	0.99±0.01	0.99±0.01	0.99±0.01
inter-N	198	159	8	31	0.98±0.01	0.98±0.01	0.97±0.03	0.98±0.03	0.98±0.03	0.98±0.03	0.98±0.03
N-ario	186	149	8	29	0.69±0.03	0.85±0.06	0.62±0.07	0.87±0.03	0.96±0.03	0.98±0.01	0.98±0.01
des-N-ar	184	148	8	28	0.61±0.13	0.84±0.08	0.58±0.08	0.82±0.05	0.89±0.06	0.91±0.06	0.91±0.06
N-udo	179	144	7	28	0.68±0.06	0.91±0.04	0.63±0.08	0.92±0.05	0.96±0.03	0.97±0.02	0.97±0.02
N'-ico	176	141	7	28	0.55±0.03	0.63±0.03	0.59±0.11	0.69±0.10	0.75±0.08	0.78±0.08	0.78±0.08
semi-N	174	140	7	27	0.99±0.02	0.99±0.02	0.99±0.01	0.99±0.01	0.99±0.01	0.99±0.01	0.99±0.01
N-ado	171	137	7	27	0.64±0.12	0.82±0.07	0.59±0.07	0.83±0.08	0.91±0.05	0.94±0.04	0.94±0.04
N-ia	164	132	7	25	0.83±0.08	0.88±0.08	0.86±0.05	0.93±0.04	0.94±0.04	0.95±0.04	0.95±0.04
N-ino	156	125	7	24	0.48±0.08	0.61±0.09	0.39±0.09	0.55±0.15	0.64±0.12	0.70±0.11	0.70±0.11
trans-N	154	124	6	24	0.93±0.02	0.93±0.02	0.90±0.08	0.93±0.07	0.94±0.07	0.94±0.07	0.94±0.07
N-aje	150	120	6	24	0.58±0.12	0.79±0.09	0.61±0.10	0.82±0.04	0.90±0.04	0.91±0.03	0.91±0.03
N-ez	147	118	6	23	0.88±0.06	0.90±0.05	0.90±0.05	0.97±0.03	0.98±0.02	0.98±0.02	0.98±0.02
N-ina	137	110	6	21	0.57±0.10	0.79±0.10	0.56±0.10	0.77±0.11	0.82±0.07	0.84±0.07	0.84±0.07
N-ita	128	103	5	20	0.47±0.15	0.62±0.11	0.52±0.08	0.71±0.06	0.78±0.06	0.83±0.05	0.83±0.05
poli-N	122	98	5	19	0.98±0.02	0.98±0.02	0.98±0.02	0.99±0.01	0.99±0.01	0.99±0.01	0.99±0.01
N-ín	118	95	5	18	0.71±0.10	0.88±0.08	0.70±0.13	0.88±0.07	0.95±0.03	0.96±0.04	0.96±0.04
multi-N	118	95	5	18	0.95±0.04	0.95±0.04	0.98±0.03	0.99±0.02	0.99±0.02	0.99±0.02	0.99±0.02
N-ano	112	90	5	17	0.41±0.06	0.58±0.09	0.46±0.10	0.62±0.09	0.70±0.09	0.80±0.04	0.80±0.04
co-N	111	89	5	17	0.98±0.02	0.98±0.02	0.96±0.06	0.97±0.06	0.99±0.02	0.99±0.02	0.99±0.02
super-N	109	88	5	16	0.97±0.04	0.97±0.04	0.97±0.03	0.98±0.03	0.98±0.02	0.98±0.02	0.98±0.02

Table 5 – continued on next page

Affix	Total	Train.	Devel.	Test	G	H	@1	@2	@3	@4	@5
N-esco	99	80	4	15	0.70±0.12	0.86±0.06	0.74±0.08	0.90±0.10	0.94±0.05	0.96±0.04	0.96±0.04
en-N-ecer	98	79	4	15	0.65±0.09	0.80±0.09	0.60±0.12	0.78±0.08	0.82±0.08	0.87±0.09	0.87±0.09
N-ato	97	78	4	15	0.60±0.09	0.73±0.11	0.59±0.17	0.71±0.17	0.84±0.13	0.88±0.08	0.88±0.08
hiper-N	96	77	4	15	0.97±0.03	0.97±0.03	0.99±0.02	0.99±0.02	0.99±0.02	0.99±0.02	0.99±0.02
V-dizo	94	76	4	14	0.97±0.03	0.99±0.02	0.97±0.03	0.99±0.02	0.99±0.02	0.99±0.02	0.99±0.02
a-N-ado	91	73	4	14	0.70±0.09	0.87±0.05	0.64±0.12	0.82±0.06	0.88±0.06	0.91±0.04	0.91±0.04
N-áceo	91	73	4	14	0.46±0.14	0.85±0.10	0.50±0.14	0.85±0.10	0.91±0.07	0.93±0.05	0.93±0.05
bi-N	91	73	4	14	1.00±0.00	1.00±0.00	0.97±0.04	0.97±0.04	0.98±0.03	0.98±0.03	0.98±0.03
N-ete	89	72	4	13	0.80±0.09	0.86±0.07	0.80±0.09	0.90±0.06	0.95±0.04	0.96±0.04	0.96±0.04
V-aje	88	71	4	13	0.93±0.08	0.93±0.08	0.93±0.06	0.97±0.03	0.99±0.02	0.99±0.02	0.99±0.02
macro-N	88	71	4	13	0.94±0.05	0.94±0.05	0.93±0.05	0.93±0.05	0.93±0.05	0.93±0.05	0.93±0.05
entre-N	87	70	4	13	1.00±0.00	1.00±0.00	0.99±0.02	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
N-ote	84	68	4	12	0.76±0.14	0.94±0.07	0.75±0.11	0.93±0.09	0.99±0.02	1.00±0.00	1.00±0.00
pos-N	84	68	4	12	0.93±0.03	0.93±0.03	0.97±0.05	0.97±0.05	0.98±0.05	0.98±0.05	0.98±0.05
V-nza	83	67	4	12	0.97±0.03	0.97±0.03	0.88±0.06	0.91±0.06	0.93±0.07	0.93±0.07	0.93±0.07
N-ificar	81	65	4	12	0.56±0.14	0.77±0.12	0.58±0.13	0.75±0.12	0.82±0.12	0.86±0.10	0.86±0.10
mini-N	80	64	4	12	0.95±0.04	0.95±0.04	0.95±0.08	0.97±0.08	0.98±0.05	0.98±0.05	0.98±0.05
intra-N	71	57	3	11	0.91±0.11	0.92±0.11	0.89±0.07	0.89±0.07	0.92±0.07	0.92±0.07	0.92±0.07
N-izo	69	56	3	10	0.88±0.07	0.93±0.07	0.67±0.08	0.81±0.06	0.82±0.06	0.83±0.04	0.83±0.04
N-oide	68	55	3	10	0.59±0.14	0.84±0.11	0.60±0.16	0.78±0.17	0.83±0.12	0.86±0.10	0.86±0.10
N'-eo	66	53	3	10	0.50±0.21	0.52±0.21	0.37±0.13	0.52±0.13	0.64±0.18	0.67±0.19	0.67±0.19
N-iento	66	53	3	10	0.54±0.13	0.81±0.07	0.54±0.13	0.73±0.18	0.90±0.11	0.92±0.09	0.92±0.09
N-ístico	66	53	3	10	0.51±0.22	0.58±0.25	0.53±0.23	0.64±0.25	0.70±0.27	0.73±0.29	0.73±0.29
auto-V	64	52	3	9	0.98±0.03	0.98±0.03	0.97±0.04	0.98±0.03	1.00±0.00	1.00±0.00	1.00±0.00
extra-N	63	51	3	9	1.00±0.00	1.00±0.00	0.96±0.05	0.98±0.03	1.00±0.00	1.00±0.00	1.00±0.00
ex-N	62	50	3	9	0.98±0.03	0.98±0.03	0.97±0.04	0.97±0.04	0.97±0.04	0.97±0.04	0.97±0.04
N-aco	61	49	3	9	0.65±0.17	0.79±0.11	0.59±0.15	0.90±0.08	0.97±0.04	0.98±0.03	0.98±0.03
hipo-N	61	49	3	9	1.00±0.00	1.00±0.00	0.98±0.03	0.98±0.03	0.98±0.03	0.98±0.03	0.98±0.03
mono-N	60	48	3	9	0.90±0.08	0.90±0.08	0.87±0.10	0.91±0.10	0.92±0.07	0.92±0.07	0.92±0.07
N-azgo	59	48	3	8	0.52±0.21	0.76±0.14	0.70±0.12	0.80±0.09	0.90±0.05	0.90±0.05	0.90±0.05
pre-V	59	48	3	8	0.98±0.04	0.98±0.04	0.97±0.05	0.97±0.05	0.97±0.05	0.97±0.05	0.97±0.05
cuasi-N	58	47	3	8	0.97±0.05	0.97±0.05	0.97±0.08	0.97±0.08	0.97±0.08	0.97±0.08	0.97±0.08
sobre-V	58	47	3	8	1.00±0.00	1.00±0.00	0.97±0.05	0.97±0.05	0.98±0.04	0.98±0.04	0.98±0.04
N-eza	56	45	3	8	0.97±0.05	0.97±0.05	0.86±0.14	0.96±0.06	1.00±0.00	1.00±0.00	1.00±0.00
pro-N	56	45	3	8	0.98±0.04	0.98±0.04	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
trans-V	56	45	3	8	0.93±0.08	0.93±0.08	0.86±0.18	0.87±0.16	0.87±0.16	0.87±0.16	0.87±0.16
ante-N	55	44	3	8	1.00±0.00	1.00±0.00	0.98±0.04	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
N-il	53	43	2	8	0.69±0.15	0.79±0.11	0.77±0.10	0.86±0.07	0.87±0.08	0.90±0.08	0.90±0.08
ultra-N	53	43	2	8	1.00±0.00	1.00±0.00	0.96±0.06	0.97±0.05	0.98±0.04	0.98±0.04	0.98±0.04
N-eño	52	42	2	8	0.67±0.11	0.90±0.08	0.64±0.22	0.77±0.20	0.87±0.10	0.89±0.11	0.89±0.11
con-V	51	41	2	8	0.93±0.06	0.93±0.06	0.97±0.05	0.97±0.05	0.97±0.05	0.97±0.05	0.97±0.05
N-ecer	48	39	2	7	0.73±0.18	0.81±0.17	0.66±0.20	0.77±0.10	0.87±0.04	0.90±0.07	0.90±0.07

Table 5 – continued on next page

Affix	Total	Train.	Devel.	Test	G	H	@1	@2	@3	@4	@5
inter-V	48	39	2	7	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
in-V	46	37	2	7	1.00±0.00	1.00±0.00	0.98±0.04	0.98±0.04	0.98±0.04	0.98±0.04	0.98±0.04
N-ota	45	36	2	7	0.77±0.12	0.90±0.10	0.74±0.17	0.88±0.12	0.95±0.07	0.95±0.07	0.95±0.07
mal-V	44	36	2	6	0.96±0.07	0.96±0.07	0.94±0.11	0.96±0.10	0.98±0.05	0.98±0.05	0.98±0.05
N-osis	43	35	2	6	0.59±0.22	0.79±0.21	0.53±0.18	0.83±0.08	0.88±0.08	0.92±0.08	0.92±0.08
N-uela	43	35	2	6	0.88±0.27	0.88±0.27	0.98±0.05	0.98±0.05	0.98±0.05	0.98±0.05	0.98±0.05
N-ería	42	34	2	6	0.77±0.14	0.88±0.11	0.72±0.14	0.86±0.11	0.94±0.08	0.96±0.07	0.96±0.07
N-ajo	42	34	2	6	0.55±0.18	0.77±0.18	0.51±0.13	0.74±0.08	0.81±0.09	0.81±0.09	0.81±0.09
co-V	42	34	2	6	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
dis-N	42	34	2	6	0.98±0.05	0.98±0.05	0.98±0.05	0.98±0.05	0.98±0.05	1.00±0.00	1.00±0.00
a-N-ear	41	33	2	6	0.59±0.14	0.74±0.16	0.53±0.16	0.68±0.12	0.75±0.08	0.79±0.07	0.79±0.07
N-iza	41	33	2	6	0.75±0.11	0.88±0.11	0.64±0.19	0.81±0.15	0.86±0.11	0.90±0.08	0.90±0.08
con-N	41	33	2	6	0.98±0.05	1.00±0.00	0.86±0.11	0.90±0.08	0.90±0.08	0.92±0.08	0.92±0.08
a-V	40	32	2	6	0.81±0.09	0.81±0.09	0.85±0.09	0.88±0.08	0.88±0.08	0.88±0.08	0.88±0.08
N-avo	38	31	2	5	0.82±0.21	0.88±0.20	0.62±0.27	0.80±0.17	0.93±0.14	0.93±0.14	0.93±0.14
peri-N	38	31	2	5	0.93±0.09	0.93±0.09	0.93±0.09	0.95±0.08	0.95±0.08	0.95±0.08	0.95±0.08
N-aza	37	30	2	5	0.55±0.26	0.71±0.17	0.44±0.16	0.68±0.24	0.82±0.21	0.91±0.14	0.91±0.14
N-uno	36	29	2	5	0.73±0.17	0.86±0.14	0.57±0.18	0.80±0.10	0.93±0.09	0.95±0.08	0.95±0.08
N-itud	34	28	2	4	0.77±0.15	0.77±0.15	0.80±0.16	0.88±0.13	0.94±0.11	0.94±0.11	0.94±0.11
archi-N	34	28	2	4	0.94±0.11	0.97±0.08	0.91±0.17	0.94±0.11	0.94±0.11	0.94±0.11	0.94±0.11
entre-V	34	28	2	4	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
N-dura	33	27	2	4	0.72±0.23	0.83±0.21	0.58±0.21	0.75±0.17	0.88±0.18	0.88±0.18	0.88±0.18
contra-V	33	27	2	4	0.88±0.13	0.88±0.13	0.91±0.12	0.97±0.08	0.97±0.08	0.97±0.08	0.97±0.08
sub-V	33	27	2	4	0.97±0.08	0.97±0.08	0.94±0.11	0.94±0.11	0.94±0.11	0.94±0.11	0.94±0.11
N-eda	30	24	2	4	0.63±0.22	0.83±0.17	0.69±0.16	0.80±0.24	0.88±0.18	0.94±0.11	0.94±0.11
tri-N	30	24	2	4	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
N-uco	29	24	1	4	0.83±0.12	0.86±0.13	0.83±0.17	0.88±0.13	0.94±0.11	0.97±0.08	0.97±0.08
N-ata	29	24	1	4	0.30±0.16	0.44±0.11	0.19±0.20	0.30±0.20	0.44±0.20	0.52±0.19	0.52±0.19
pluri-N	28	23	1	4	0.97±0.08	0.97±0.08	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
a-N	26	21	1	4	0.88±0.13	0.88±0.13	0.86±0.18	0.94±0.11	0.94±0.11	0.97±0.08	0.97±0.08
V-ata	26	21	1	4	0.77±0.19	0.77±0.19	0.91±0.12	0.91±0.12	0.91±0.12	0.91±0.12	0.91±0.12
vice-N	25	20	1	4	0.94±0.11	0.94±0.11	0.91±0.12	0.91±0.12	0.94±0.11	0.97±0.08	0.97±0.08
supra-N	24	20	1	3	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
N-eno	23	19	1	3	0.52±0.24	0.77±0.23	0.59±0.32	0.70±0.35	0.81±0.24	0.89±0.16	0.89±0.16
N-ío	23	19	1	3	0.33±0.23	0.74±0.27	0.40±0.27	0.77±0.16	0.81±0.17	0.81±0.17	0.81±0.17
N-uelo	23	19	1	3	0.77±0.23	0.89±0.16	0.89±0.16	0.89±0.16	0.92±0.14	0.92±0.14	0.92±0.14
V-toria	22	18	1	3	0.92±0.14	0.92±0.14	0.74±0.43	0.77±0.44	0.77±0.44	0.77±0.44	0.77±0.44
cuatri-N	22	18	1	3	1.00±0.00	1.00±0.00	0.96±0.10	0.96±0.10	0.96±0.10	0.96±0.10	0.96±0.10
N-ucho	21	17	1	3	0.59±0.32	0.70±0.30	0.81±0.17	0.92±0.14	1.00±0.00	1.00±0.00	1.00±0.00
V-e	21	17	1	3	0.29±0.20	0.29±0.20	0.40±0.36	0.44±0.33	0.44±0.33	0.48±0.33	0.48±0.33
para-N	21	17	1	3	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
a-N-ecer	20	16	1	3	0.37±0.26	0.51±0.24	0.37±0.26	0.59±0.22	0.70±0.20	0.74±0.22	0.74±0.22

Table 5 – continued on next page

Affix	Total	Train.	Devel.	Test	G	H	@1	@2	@3	@4	@5
N-edo	20	16	1	3	0.51±0.29	0.92±0.14	0.55±0.37	0.96±0.10	0.96±0.10	1.00±0.00	1.00±0.00
N-tivo	20	16	1	3	0.44±0.40	0.44±0.40	0.44±0.17	0.55±0.23	0.74±0.22	0.77±0.23	0.77±0.23
N-iego	20	16	1	3	0.48±0.17	0.59±0.27	0.59±0.32	0.74±0.22	0.85±0.17	0.85±0.17	0.85±0.17
N-ucha	20	16	1	3	0.92±0.14	0.92±0.14	0.92±0.14	0.96±0.10	1.00±0.00	1.00±0.00	1.00±0.00
N-ing	20	16	1	3	0.85±0.33	0.92±0.14	0.89±0.16	0.96±0.10	0.96±0.10	1.00±0.00	1.00±0.00
N-oteca	20	16	1	3	0.74±0.27	0.89±0.16	0.44±0.33	0.77±0.28	0.92±0.14	0.92±0.14	0.92±0.14
en-N-izar	20	16	1	3	0.22±0.23	0.37±0.35	0.48±0.33	0.66±0.33	0.66±0.33	0.77±0.28	0.77±0.28
hexa-N	20	16	1	3	0.89±0.16	0.96±0.10	0.89±0.16	0.96±0.10	0.96±0.10	0.96±0.10	0.96±0.10
V-ndo	19	16	1	2	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
V-ato	19	16	1	2	0.55±0.39	0.55±0.39	0.72±0.36	0.72±0.36	0.72±0.36	0.72±0.36	0.72±0.36
bien-N	19	16	1	2	0.83±0.25	0.83±0.25	0.94±0.16	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
N-érrimo	18	15	1	2	0.50±0.35	0.55±0.39	0.61±0.33	0.66±0.35	0.77±0.26	0.77±0.26	0.77±0.26
ciber-N	18	15	1	2	0.94±0.16	0.94±0.16	0.94±0.16	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
uni-N	18	15	1	2	0.94±0.16	0.94±0.16	0.94±0.16	0.94±0.16	0.94±0.16	0.94±0.16	0.94±0.16
hetero-N	17	14	1	2	1.00±0.00	1.00±0.00	0.83±0.35	0.83±0.35	0.83±0.35	0.83±0.35	0.83±0.35
N-íneo	16	13	1	2	0.55±0.16	0.66±0.25	0.61±0.41	0.94±0.16	0.94±0.16	0.94±0.16	0.94±0.16
N-acho	16	13	1	2	0.88±0.33	0.88±0.33	0.88±0.22	0.94±0.16	1.00±0.00	1.00±0.00	1.00±0.00
N-itar	16	13	1	2	0.55±0.39	0.61±0.33	0.50±0.25	0.66±0.25	0.77±0.26	0.77±0.26	0.77±0.26
V-orio	16	13	1	2	0.66±0.25	0.77±0.26	0.61±0.33	0.66±0.25	0.94±0.16	0.94±0.16	0.94±0.16
N-eto	16	13	1	2	0.38±0.22	0.55±0.30	0.44±0.39	0.55±0.39	0.77±0.26	0.77±0.26	0.77±0.26
V-ido2	16	13	1	2	0.88±0.22	0.88±0.22	0.83±0.25	0.88±0.22	0.88±0.22	0.88±0.22	0.88±0.22
N-ífico	16	13	1	2	0.55±0.30	0.83±0.25	0.44±0.39	0.72±0.26	0.72±0.26	0.77±0.26	0.77±0.26
en-V	16	13	1	2	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
hemi-N	16	13	1	2	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
N-ingo	15	12	1	2	1.00±0.00	1.00±0.00	0.88±0.22	0.88±0.22	0.94±0.16	1.00±0.00	1.00±0.00
V-oso	15	12	1	2	0.61±0.41	0.61±0.41	0.72±0.36	0.83±0.25	0.88±0.22	0.88±0.22	0.88±0.22
arqueo-N	15	12	1	2	1.00±0.00	1.00±0.00	0.83±0.35	0.83±0.35	0.88±0.33	0.88±0.33	0.88±0.33
cis-N	15	12	1	2	0.88±0.22	0.88±0.22	0.77±0.44	0.88±0.33	0.94±0.16	0.94±0.16	0.94±0.16
deca-N	15	12	1	2	0.50±0.25	0.50±0.25	0.50±0.35	0.50±0.35	0.55±0.39	0.55±0.39	0.55±0.39
dis-V	15	12	1	2	0.94±0.16	0.94±0.16	0.94±0.16	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
endo-N	15	12	1	2	0.77±0.26	0.77±0.26	0.94±0.16	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
super-V	15	12	1	2	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
N-orro	14	12	1	1	0.55±0.52	1.00±0.00	0.66±0.50	0.88±0.33	0.88±0.33	1.00±0.00	1.00±0.00
a-N-izar	13	11	1	1	0.44±0.52	0.44±0.52	0.33±0.50	0.33±0.50	0.33±0.50	0.44±0.52	0.44±0.52
N-és	13	11	1	1	0.44±0.52	0.55±0.52	0.55±0.52	0.77±0.44	0.77±0.44	0.77±0.44	0.77±0.44
V-io	13	11	1	1	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33
V-era	13	11	1	1	1.00±0.00	1.00±0.00	0.77±0.44	0.88±0.33	0.88±0.33	1.00±0.00	1.00±0.00
N-amen	13	11	1	1	0.55±0.52	0.88±0.33	0.66±0.50	0.88±0.33	1.00±0.00	1.00±0.00	1.00±0.00
V-nda	13	11	1	1	0.77±0.44	0.77±0.44	0.88±0.33	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
ex-V	13	11	1	1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
exo-N	13	11	1	1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
per-V	13	11	1	1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00

Table 5 – continued on next page

Affix	Total	Train.	Devel.	Test	G	H	@1	@2	@3	@4	@5
trans-N-ar	13	11	1	1	0.55±0.52	0.66±0.50	0.88±0.33	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
N-icio	12	10	1	1	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33
N-dumbre	12	10	1	1	0.11±0.33	0.11±0.33	0.22±0.44	0.44±0.52	0.66±0.50	0.66±0.50	0.66±0.50
circun-N	12	10	1	1	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33
epi-N	12	10	1	1	0.88±0.33	0.88±0.33	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
giga-N	12	10	1	1	1.00±0.00	1.00±0.00	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33
iso-N	12	10	1	1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
tetra-N	12	10	1	1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
V-azo	11	9	1	1	0.88±0.33	0.88±0.33	0.88±0.33	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
N-nte	11	9	1	1	0.44±0.52	0.77±0.44	0.55±0.52	0.55±0.52	0.77±0.44	0.77±0.44	0.77±0.44
N-ción	11	9	1	1	0.11±0.33	0.11±0.33	0.66±0.50	0.77±0.44	1.00±0.00	1.00±0.00	1.00±0.00
N-ólatra	11	9	1	1	0.00±0.00	0.00±0.00	0.11±0.33	0.11±0.33	0.11±0.33	0.11±0.33	0.11±0.33
bien-V	11	9	1	1	0.88±0.33	0.88±0.33	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
N-ena	11	9	1	1	0.66±0.50	0.66±0.50	0.44±0.52	0.55±0.52	0.77±0.44	0.77±0.44	0.77±0.44
com-N	11	9	1	1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
euro-N	11	9	1	1	0.88±0.33	0.88±0.33	0.88±0.33	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
puto-N	11	9	1	1	1.00±0.00	1.00±0.00	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33
N-ango	10	8	1	1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
N-aja	10	8	1	1	0.22±0.44	0.66±0.50	0.22±0.44	0.66±0.50	0.77±0.44	0.77±0.44	0.77±0.44
V-ina	10	8	1	1	0.88±0.33	0.88±0.33	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
ante-V	10	8	1	1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
N-uca	10	8	1	1	0.66±0.50	0.66±0.50	0.88±0.33	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
N-icia	10	8	1	1	0.66±0.50	0.66±0.50	0.77±0.44	0.77±0.44	0.77±0.44	0.77±0.44	0.77±0.44
N-anga	10	8	1	1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
bis-N	10	8	1	1	0.66±0.50	0.77±0.44	0.66±0.50	0.77±0.44	1.00±0.00	1.00±0.00	1.00±0.00
com-V	10	8	1	1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
crio-N	10	8	1	1	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33
demi-N	10	8	1	1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
des-N-ado	10	8	1	1	0.33±0.50	0.55±0.52	0.66±0.50	0.77±0.44	0.77±0.44	0.88±0.33	0.88±0.33
dino-N	10	8	1	1	0.55±0.52	0.55±0.52	0.88±0.33	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
eco-N	10	8	1	1	0.66±0.50	0.66±0.50	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
ecto-N	10	8	1	1	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33
ex-N-ar	10	8	1	1	0.55±0.52	0.88±0.33	0.66±0.50	0.77±0.44	0.77±0.44	0.77±0.44	0.77±0.44
geronto-N	10	8	1	1	0.88±0.33	0.88±0.33	0.88±0.33	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
hemo-N	10	8	1	1	0.66±0.50	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33	0.88±0.33
ultra-V	10	8	1	1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00

Table 5