Proceedings of the EURALI Workshop @LREC2020, pages 42–50
Marseille, 20 June 2022

Automatic Verb Classifier for Abui (AVC-abz)

František Kratochvíl, George Saad, Jiří Vomlel, Václav Kratochvíl
Department of Asian Studies, Palacký University Olomouc, Czech Republic,
Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic
{frantisek.kratochvil, george.saad}@upol.cz, {vomlel, velorex}@utia.cas.cz

Abstract

We present an automatic verb classifier system that identifies inflectional classes in Abui (AVC-abz), a Papuan language of the Timor-Alor-Pantar family. The system combines manually annotated language data (the learning set) with the output of a morphological precision grammar (corpus data). The morphological precision grammar is trained on a fully glossed smaller corpus and applied to a larger corpus. Using the k-means algorithm, the system clusters inflectional classes discovered in the learning set. In the second step, Naive Bayes algorithm assigns the verbs found in the corpus data to the best-fitting cluster. AVC-abz serves to advance and refine the grammatical analysis of Abui as well as to monitor corpus coverage and its gradual improvement.

Keywords: automatic verb classifier, endangered languages, head-marking languages, Papuan

1. Analytical problem: Abui verb classes

Across languages, verbs are known to be sensitive to their syntactic context in various ways. Their meaning may remain constant (a paraphrase) or it may differ (e.g. number and type of arguments and their role). Their distribution across varying contexts is used to identify verbal classes that capture the meaning of the verb. Levin (2015) is an excellent overview of the various approaches to verb alternations that have been developed over the last 50+ years, starting with the work on case alternations (Fillmore, 1970), through valence databases, such as FrameNet or the Proposition Bank (Baker et al., 1998; Palmer et al., 2005) to recent typological studies, such as ValPal (Hartmann et al., 2013). Such investigations are very labour-intensive and take years to complete for well-resourced language but are rarely undertaken for low-resourced languages. The workflow described here offers a significant acceleration to this endeavour by combining a learning set with the output of corpus data.

Abui is a head-marking language which records the argument configuration of the verb in its morphology; the verbal complements are indexed on the verb (their type, person and number).[1] Therefore, the morphological indexing offers a formal means of classification, where verb stems can be classified according to their indexing of arguments analogously to the dependent-marking languages where such information can be extracted from the syntactically annotated corpus and where significant advances have indeed been made. In particular, we follow the work on automatic verb classification undertaken on well-resourced languages in using an abstract feature space that is the input for mathematical clustering tools (Merlo and Stevenson, 2001; Kipper et al., 2008; Sun and Korhonen, 2009).

In the remainder of this section we give a brief overview of the Abui verbal morphology. Abui is notable for its argument realisation, which has been argued to be sensitive to semantic rather than syntactic features where the verbal stems show a low degree of lexical stipulation, i.e. the verb stems are compatible with a large number of morphological devices and their meaning is sometimes adjusted during this process (Kratochvíl, 2007; Kratochvíl, 2011). This issue has been discussed and elaborated in other publications (Fedden et al., 2014; Kratochvíl, 2014; Kratochvíl and Delpada, 2015; Saad, 2020a; Kratochvíl et al., 2021).

1.1. Abui verbal morphology and argument indexing

Verbs are at the heart of morphological complexity in Abui. Table 1 presents a schematic morphological template of the Abui verb, where the first line indicates the slot numbers, the second line the categories marked in each slot, and the subsequent lines the values attested in each slot. The table shows that (i) the root may be preceded by up to three person-number prefixes indexing various types of undergoer arguments or an incorporated noun (slots -1 to -3) and/or (ii) by the causative or applicative prefixes (slot -4).[2] Many roots mutate to distinguish two stems (perfective and imperfective) and sometimes three (+ inceptive). Roots may be followed by (iii) up to three aspectual slots (+1 to +3) and two mood slots (+4 to +5). The table records the values attested in each slot, represented here just by their glosses. For more details on the root mutation and aspectual suffixation, see (Kratochvíl et al., 2021).

[1] Abui [abz] is a Timor-Alor-Pantar language of Eastern Indonesia. Over the last almost two decades, we have collected a corpus of roughly 22,500 sentences, of which about 6,200 have been glossed (Kratochvíl, 2022). The corpus consists of various genres and includes also elicited data.

[2] According to Siewierska (2013), systems marking undergoers alone (leaving actors unmarked) are rare, constituting only about 7% of her sample. In the Alor-Pantar family, undergoer marking is a common trait.

-4	-3	-2	-1	0	+1	+2	+3	+4	+5
EXT	EXT/U_3	U_2	U_1	$root_{mutation}$	ASP_1	ASP_2	ASP_3	$MOOD_1$	$MOOD_2$
CAUS	BEN	LOC	PAT	$root_{pfv}$	INCP	INCH	STAT	PRIOR	HORT
APPL	GOAL	REC	N	$root_{ipfv}$	STAT	PFV	PROG	REAL	PROH
				$root_{incp}$		PRF			

Table 1: Morphological template of the Abui verb

1.2. Person-number prefixes

The prefixal slots of the Abui verb index objects, applicatives, and causatives. Objects are primarily indexed by a collection of five person-number prefix series, which are given in Table 2.[3] To a large extent, each series is phonologically distinct (e.g. series PAT singular prefixes tend to end in *a*); but some plural forms are syncretic (e.g. series PAT and LOC). The person-number prefixes occur in slots -3, -2, and -1, where slot -1 is reserved to series PAT series (listed in the second column) or incorporated nouns.

PERSON	PAT	REC	LOC	GOAL	BEN
1SG	*na-*	*no-*	*ne-*	*noo-*	*nee-*
2SG	*a-*	*o-*	*e-*	*oo-*	*ee-*
1PL.EXCL	*ni-*	*nu-*	*ni-*	*nuu-*	*nii-*
1PL.INCL	*pi-*	*pu-/po-*	*pi-*	*puu-/poo-*	*pii-*
2PL	*ri-*	*ro-/ru-*	*ri-*	*ruu-/roo-*	*rii-*
3	*ha-*	*ho-*	*he-*	*hoo-*	*hee-*
3.REFL	*da-*	*do-*	*de-*	*doo-*	*dee-*
DISTR	*ta-*	*to-*	*te-*	*too-*	*tee-*

Table 2: Abui person-number indexing paradigm

Some Abui verbs may combine with multiple prefix series, as shown in (1) where several prefix combinations of the verb *wik* 'carry' are listed. The PAT series-indexed form *ha-wik* in (1a) is used when an animate object, *kaai* 'dog', is involved. In (1b), the LOC series-indexed *he-wik* indexes a definite inanimate object, while the BEN series-indexed *hee-wike* involves a human benefactor for whom an object is carried (implied or contextually available). In (1c), the GOAL series-indexed *hoo-wik* is a type of causative construction, where the first person singular agent passes firewood to another person to carry; the secondary agent is indexed with the GOAL prefix. Finally, in (1d), the plain form *wik* is used when the object *sura foqa do* 'this big book' is topicalised and the information about the argument structure is contextual (i.e. the speaker assumes their responsibility for carrying the book). To sum up, the various combinations of person indexes and the verb *wik* 'carry (in arms)' distinguish object types, modify argument structure, and are sensitive to discourse structure.

(1) a. Bui kaai **ha**-wik.
name dog 3.I-carry.IPFV
'Bui is carrying her dog in her arms.' [N2011.9]

b. A-táng do mi **he**-wik,
2SG.INAL-hand PROX take 3.III-carry.IPFV
hee-wik-e!
3.V-carryIPFV-PROG
'Carry it in your hands, carry it for him!' [N2011.3]

c. Na ara mi **hoo**-wik.
1SG.AGT firewood take 3.IV-carry.IPFV
'I give him firewood to carry.' [N2011.6]

d. Sura foqa do baai wik-e?
book big PROX ADD carry.IPFV-PROG
'This big book too, (should I) carry?' [EVY.1238]

The person-number combinations of the verb *wik* 'carry' are not generalisable to other verbs and as we will show in section 2.1, verbs show various gaps in their compatibility with the affixes.

When more than one person-number prefixes co-occur, the patientive prefix (PAT) must occur in slot -1 (U_1). This is shown in (2), where the verb *minang* 'remember' combines with two person-number prefixes in slots -1 (PAT) indexing the experiencer and in slot -2 (LOC) indexing the medicine (stimulus).

(2) Ata di he-daweng
name 3.AGT 3.AL-medicine
he-da-minang-di.
3.LOC-3.REFL.PAT-remember-INCH
'Ata remembered his medicine.'

1.3. Light verbs

The second part of the argument-marking system consists of a set of light verbs which attach before the lexical verb and may take their own person-number indexes. The main function of the light verbs is to adjust the valency of the verb in a manner similar to adpositions (prepositions and postpositions) in other languages.[4] Table 3 lists the most common light verbs that modify the argument frame of the main verb by highlighting human objects or by adding human goals or companions. All five light verbs are used to refine the marking of human participants affected by the event described by the main verb and have been analysed as differential argument marking devices (Kratochvíl, 2014).

[3] The DISTR prefixes index reciprocals and distributives.

[4] Some of the multiple prefix series in Alor-Pantar languages (including the Abui prefixes in Table 3) have their origin in complex verbs (Klamer and Kratochvíl, 2018).

category	morphology
human undergoer	U.V-**GIVE**=main.verb
experiencer	U.IV-**INSIDE**=main.verb
human goal (proximal)	U.IV-**TOUCH**=main.verb
human goal (remote)	U.IV-**THROW**=main.verb
companion	U.I-**JOIN**=main.verb

Table 3: Abui light verbs attested in complex verbs

An example of the light verb use is shown in (3), where the verb *he-fikang* 'guard, look after s.t.' combines with the light verb clitic *hee-l*. The light verb differentiates the human object *ama* 'person'. The meaning of the main verb *he-fikang* 'guard, look after s.t.' shifts due to the light verb *hee-l*= addition to 'respect s.o., pay attention to s.o.'.

(3) Deri di ama
name 3.AGT person
hee-l=he-fikang, hare
3.BEN-GIVE=3.LOC-respect.IPFV so
do-wa tanga naha
3REFL.REC-participate speak.IPFV not
'Deri respects people, so she does not talk (much)' [NB9.103]

1.4. Applicative and causative prefixes

The third part of the argument marking system is formed by applicative prefixes (*lang-*, *ming-*) and the causative prefix *ong-* (attaching in slot -4) which extend the valency of the verb but do not index number or gender features of the added arguments. An example of the applicative prefix *lang-* can be seen in (4) where the reduplicated verb *mara* 'go up, climb' combines with *lang-* to add the nominal *dieng-pe* 'kitchen' to the argument structure of the intransitive motion verb.

(4) kaai de-tamai dieng-pe
dog 3.REFL.III-keep.doing.IPFV kitchen
lang-mara∼mara
APPL-RED∼go.up.IPFV
'The dog is entering kitchen all the time.' [EBD.047]

Complex predicates consisting of a light verb and a main verb are also compatible with applicatives, as shown in 5. The complex verb *na-da=sama* 'be with me' combines with the applicative *ming-* to include the time description into the argument structure of the verb.

(5) tung-ai loohu **ming-na-da**=sama
year-root be.long APPL-1SG.I-JOIN=be.with
'may I have a long life!' (lit. 'may long years be with me!') [EBD.7.15.7c]

The above examples illustrate that undergoer prefix series, light verbs, and applicative/causative prefixes constitute a complex argument marking system. Through detailed examination of the combinatorics of verb stems and the undergoer-marking material we can arrive at a classification of verb stem and at a better understanding of the semantic contribution of the prefixes. In the following sections we will describe the workflow that we have designed for this purpose.

2. System design

The Automatic Verb Classifier for Abui processes two types of data through three analytical modules. The data constituting the *learning set* is manually compiled and its structure is described in section 2.1. The *learning set* is analysed with the k-means clustering technique which groups the data into a pre-specified number of clusters, as described further in section 2.2.
The next step is to include the pre-processed *corpus data* and examine the fit with the k-means based clustering. The pre-processing of the *corpus data* was described in a separate publication (Zamaraeva et al., 2017). We offer a brief summary of this work in section 2.3. In section 2.4 we describe how the comparison of the clustering based on the learning set is implemented for the *corpus data*. The final part of the module will be a Bayesian network analysis whose intended purpose is briefly discussed in section 3. The workflow of the AVC-abz system is visualised in Figure 1. It processes a manually curated and complete training data to train a classifier system whose outputs (clustering, visualisation, membership lists) aid the linguistic analysis. The Bayesian Network Analysis Module is not yet finished.

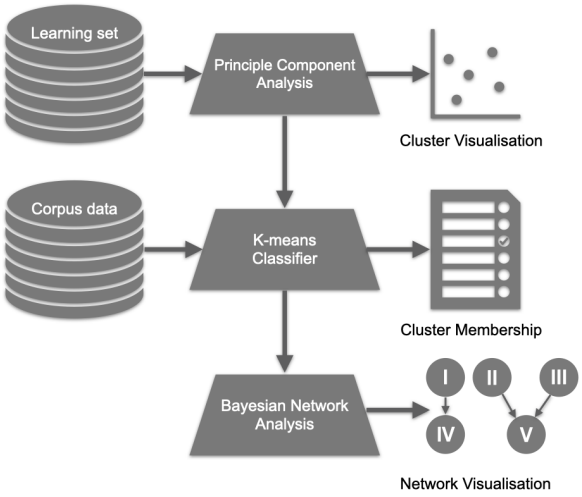

Figure 1: The components of the Automatic verb classifier for Abui (AVC-abz)

The system is implemented in an online interface available at `http://gogo.utia.cas.cz/abui/`. The four data sets (2 learning sets, 2 corpus data sets) and the source code in R are available at `https://github.com/fanacek/AVC-abz`.

2.1. Learning set structure

The *learning set* is a manually built database that maps the inflectional profiles of selected verb stems. It was

Form	Gloss	Feature
Ø-	stem alone	A
Ca-	patient (PAT)	B
Ce-	location (LOC)	C
Cee-	benefactive (BEN)	D
Co-	recipient (REC)	E
Coo-	goal (GOAL)	F
Cee-l=	human undergoer	G
Coo-q=	animate goal I	H
Coo-pang=	animate goal II	I
Ca-da=	companion (JOIN)	J
Coo-mi=	inward goal	K
ming-	applicative I (APPL)	L
lang-	applicative II (APPL)	M
ong-	causative (CAUS)	N

Table 4: Inflectional features in the learning sets

designed by the authors in close collaboration with a team of Abui speakers who are responsible for the accuracy of the grammatical information.[5]

Currently, the *AVC-abz* interface includes two learning sets counting 150 and 356 verb profiles respectively, tracking the compatibility of the verbs with 26 inflectional features. Table 4 lists fourteen of these features. The values of seven features are exemplified for six verbs in Table 5.

Table 4 lists possible morphological features of the Abui verb. Feature A is stem attested bare. Features B-F refer to person-number prefixes listed in Table 2, where the *C-* symbol is used as a shorthand for the various consonants distinguishing the person-number.[6] Features G-K combine light verbs and person-number prefixes, as listed in Table 3. Finally, forms L-K are applicative and causative prefixes. The morphological forms are accompanied by a short semantic characterisation (and a gloss).[7]

In Table 5, we exemplify the structure of the learning set (columns listing conditions A-G only). The full dataset can be viewed using the *Data* tab of the *AVC-abz* interface.

[5]While we are aware of the typical drawbacks of elicitation data that relies on a small number of speakers, such as accidental mistakes, gaps, or false negatives (forms that may sound unnatural in isolation may be fine in natural speech), we continue expanding the learning set, refining it with new verbs and information. Given the low-resourced status of Abui, it is unreasonable to expect that the corpus will reach the size where we could rely on it alone as a source of morphological information.

[6]For the ease of exposition we ignore the plural forms here. The characteristic vowel patterns in singular appear to have much higher frequency in the corpus anyway.

[7]While the gloss labels are suggestive of semantic roles, their exact semantic contribution is more complex and ultimately one of the puzzles we are working towards solving. The labels should therefore be interpreted as preliminary place-holders.

Stem	A	B	C	D	E	F	G
wik 'carry'	+	+	+	+	+	+	+
fanga 'say'	+	+	+	+	+	+	+
aquta 'blind'	+	+	-	+	+	+	+
took 'pour'	+	-	+	+	+	-	-
yaa 'go'	+	-	+	+	+	+	-
bai 'angry'	-	-	+	-	-	-	-

Table 5: Examples of Abui verb feature profiles

2.2. Principle component analysis and cluster visualisation

Using automatic clustering methods we find clusters of verbs that share a similar feature profile, i.e. the verbs occur in the same or similar morphological environments. We use the k-means clustering technique, which groups the data into a pre-specified number of clusters (k), while minimising the inter-cluster distance (d), which is defined as the total sum-of-squares distance of cluster members to the cluster mean.

The optimal number of clusters is not known in advance. The k-means technique allows us to employ expert human judgement and experiment with the optimal cluster number. To aid the human judgment we use a visualisation technique that plots the inter-cluster distance for each number of clusters. The idea here is that typically there is an elbow-like shape in the plot that enables us to identify the optimal number of clusters. The threshold for the optimal number for clusters is set so that the inter-cluster distance (d) decreases slowly after the threshold.

An example of the inter-cluster distance plot is given in Figure 2 where the plot shows an elbow-like dip in the inter-cluster distance. The inter-cluster distance decreases rapidly until 24 clusters, but starts to decrease gradually after 25 clusters. Therefore we set the threshold at 25 clusters.

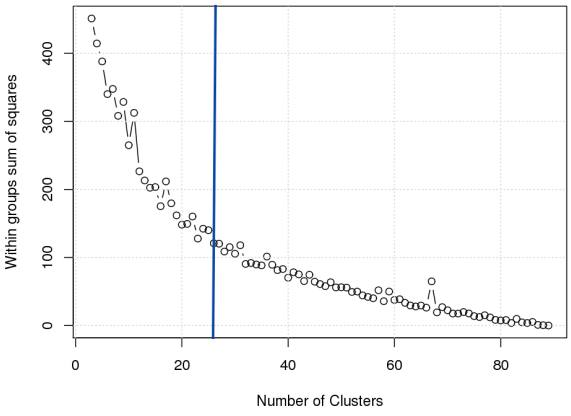

Figure 2: The inter-cluster distance plot for the learning set *Abui verbs (356) v. 2020.*

For comparison, when we examine an older and smaller learning set of 150 verbs, we can see in the inter-cluster distance plot, shown in Figure 3, that the

threshold value is lower. The inter-cluster distance decreases rapidly until 17 clusters, but starts to decrease gradually after 18 clusters but the characteristic elbow-shape is less obvious.
We conjecture that the size of the learning set influences still the threshold value and therefore the learning set should be further expanded with new verbs until the threshold remains stable. This is a very useful information to guide the laborious construction of the learning set which requires a lot of time of highly-trained native speakers.

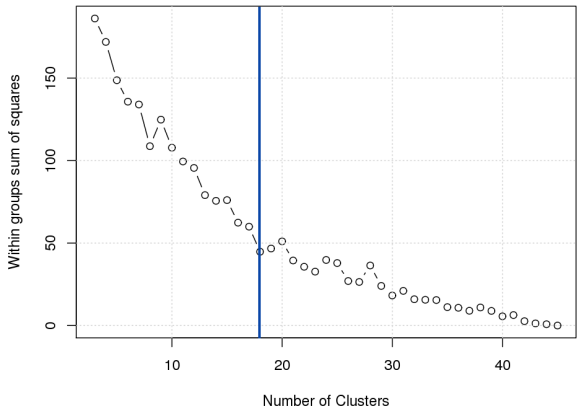

Figure 3: The inter-cluster distance plot for the learning set *Abui verbs (150) v. 2020*.

We have built a visual interface, which allows us to examine the clusters in 3D space using the first three principal components as vectors. The visual interface represents each verb by a sphere. In case there are more than one verbs in one position, the diameter of the sphere is increased proportionally to the number of verbs in that position.[8]
The interface allows the 3D space to be rotated to examine whether the cluster members are not too far apart allowing a visual inspection of the clusters. Thanks to the rotation feature, we can see the shape of the cluster, for example, whether its members are aligned along a line, lie within a sphere, are scattered far apart without an obvious geometric relation, etc.
A screenshot of the visualisation of the *Abui verbs (356) v. 2020* learning set can be seen in Figure 4. The number of clusters is set here at 25 and clusters are numbered and coloured.
The content of each cluster is listed in a separate window called *Cluster Membership* under the same number as used in the 3D plot. We use the data point closest to the mean value to characterise each cluster; i.e. the verb nearest to the mean value becomes the cluster label in the *Cluster membership* window (see section 2.4).
The cluster visualisation interface includes a version control, so that we can load newer versions of learning sets and check the differences in analysis. Similarly, the gradually improving coverage of our corpus data is also stored, as described in the next section.

[8]Please note that in Figure 4 the largest sphere contains 164 verbs.

2.3. Corpus data harvesting

Presently, the Abui corpus is managed using the SIL Toolbox (SIL International, 2015) and SIL Fieldworks (SIL International, 2017). Both tools support simple concordance functions but lack more powerful distributional analysis tools needed to tackle the present problem. We therefore rely on a workflow, described in (Zamaraeva et al., 2017), where a morphological grammar of Abui is inferred from interlinear glossed text (IGT) extracted from the glossed part of the Abui corpus and applied on the entire corpus following a workflow described in (Bender et al., 2014).
The workflow is built on the precision grammar architecture known as the Grammar Matrix project (Bender et al., 2002; Bender et al., 2010) which supports the creation of starter-kit precision grammars on the basis of the lexical and typological language profile and allows for specification of position classes and lexical rules (O'Hara, 2008; Goodman, 2013). In addition, the precision grammar is enhanced by the information retrieved from existing collections of IGT, using methods developed by Lewis and Xia (2008) and Georgi (2016). The system is implemented as Matrix-ODIN Morphology or 'MOM' (Wax, 2014; Zamaraeva, 2016). It extracts from a corpus of IGT information such as (i) sets of affixes grouped in position classes; (ii) for each affix also its gloss; (iii) input for each position class. The MOM system can generate a feature matrix of the same structure as the learning set.
Affixes can be expressed as a graph whose nodes represent the input relations. This can be illustrated using the example sentence in (6).

(6) he-ha-luol tila bataa ha-tang
3.LOC-3.PAT-follow rope tree 3.PAT-branch
he-tilaka mai neng nuku di mii
3.LOC-hang REAL man one 3.AGT take.PFV
ya ho-puna ba natea.
SEQ 3.REC-hold.IPFV SIM stand
'in the next one, there was a rope hanging on the tree branch when a man came and took it and remained standing there holding it.'

Classifying verbal stems is approached as a co-occurrence problem: given segmented and glossed IGT, we determine which stems co-occur with which types of affixes. Using the data in example (6), we see that (i) the verbs *mii* 'take' and *natea* 'stand' can occur freely (Feature I); (ii) the verb *tilaka* 'hang' with the prefix *he-* (Feature III); (iii) the verb *luol* 'follow' can combine with the prefixes *he-* and *ha-* (Feature II and III); and (iv) the verb *puna* 'hold' with the prefix *ho-* (Feature V).
The resulting co-occurrences can be explored visually, as described in (Lepp et al., 2019) or listed (as input

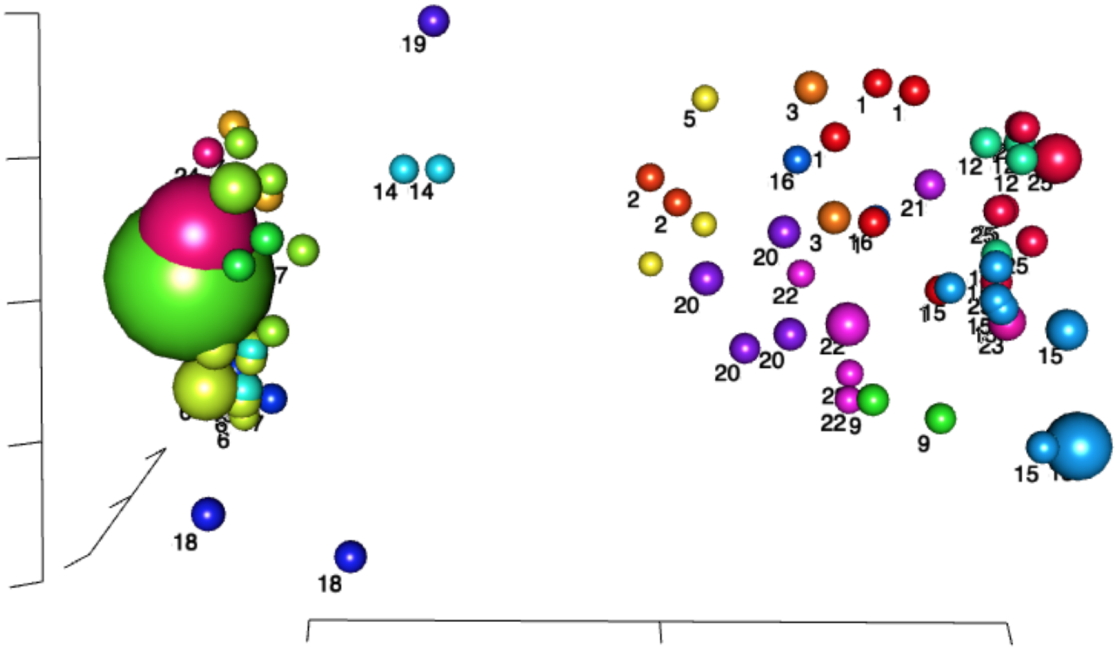

Figure 4: The visualisation of the structure of the learning set *Abui verbs (356) v. 2020.*

for the *AVC-abz*). The process is outlined in Figure 5, presented in (Zamaraeva et al., 2017), and explained briefly below.

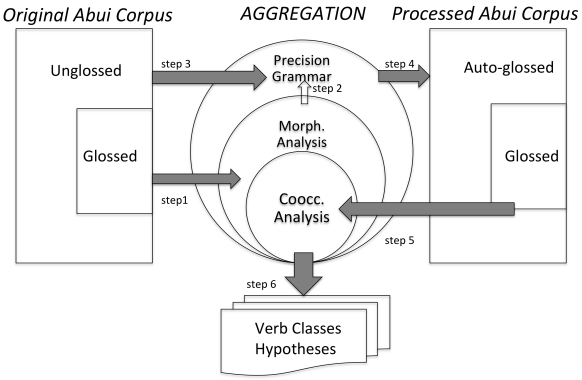

Figure 5: The components of the corpus-processing (from Zamaraeva et al., 2017).

The AGGREGATION workflow allows bootstrapping the morphological information in the glossed part of the corpus which can be used to automatically analyse words that have not been manually glossed. In this way we can process much more data and verify the validity of the clustering proposed based on the learning set, as described in section 2.2.

2.4. Cluster lists comparison

The fourth component of the system is a *Cluster list* feature, which is built on a Naive Bayes clasiffier. Naive Bayes is a simple technique that assumes independence among features. Using the *learning set* features a naive Bayes classifier considers the features of each verb in the *corpus set* and assigns each verb to the most probable cluster.

The output of the Naive Bayes is listed as a table, as shown in Figure 6. The verb stems that are assigned by the Naive Bayes to the same cluster as they were by the k-means classifier are listed in bold face. Stems found only in one of the two datasets are listed in regular plain face. Finally, stems that are assigned differently by both classifiers are listed in italics. The typographic distinction aids the human expert to quickly evaluate the classification, check the data values by clicking the listed verb stem to examine their features.

The verb profiles can be examined using the *Data* tab shown in Figure 7. Thus mismatches between the curated and automatically derived data sets do not necessarily indicate system errors. Instead, they represent cases in which corpus analysis can further our understanding of the language at hand.

Given the small size of descriptive corpora, the generated feature array will have properties of a sparse matrix, where most elements are zero, not because the combination is impossible, but because the target morpheme sequence is not attested in the corpus. The sparsity of the feature matrix is *de facto* a metric of the corpus coverage of the selected phenomena. Mathematical methods exist to solve sparse matrixes so that they can serve as a reliable input for machine learning and clustering algorithms, whose output in turn aids linguistic analysis. For example, Fisher Information evaluation (which feature(s) predicts the class membership the best) informs the fieldwork practice (i.e. which constructions should be elicited and in what sequence). The clustering analysis helps detect morphemes with similar distributional properties. The analysis can be

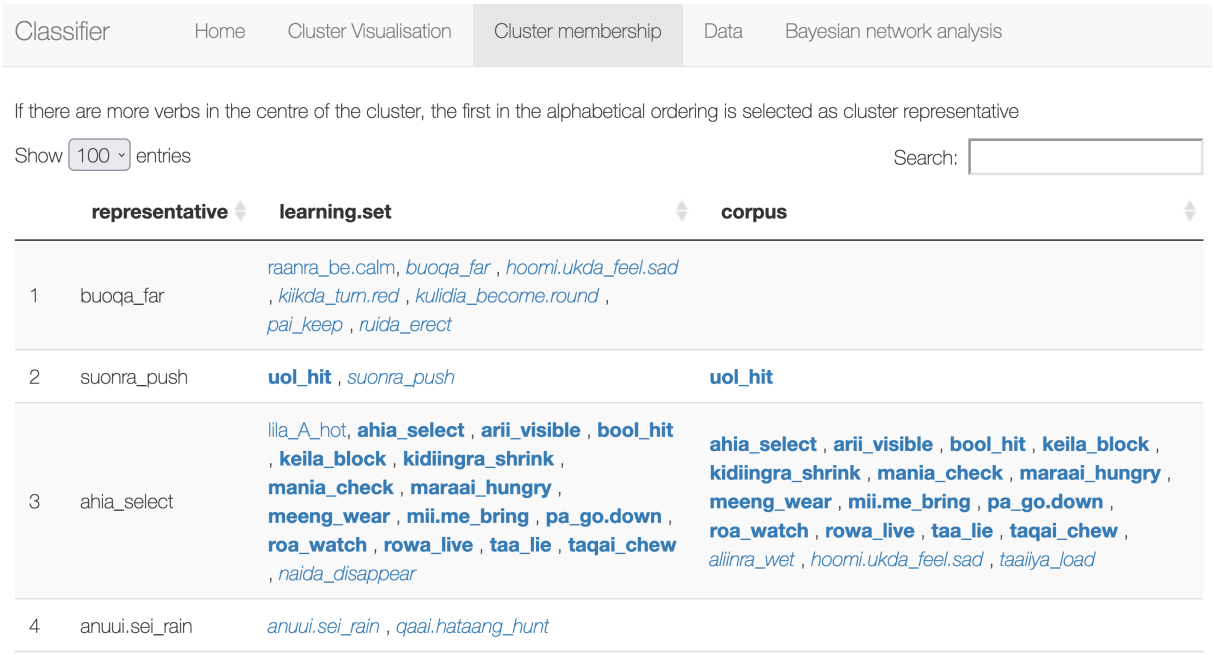

Figure 6: The cluster list comparison. The left column lists the cluster membership as suggested by the k-means algorithm. The right column is the output of a Naive Bayes algorithm.

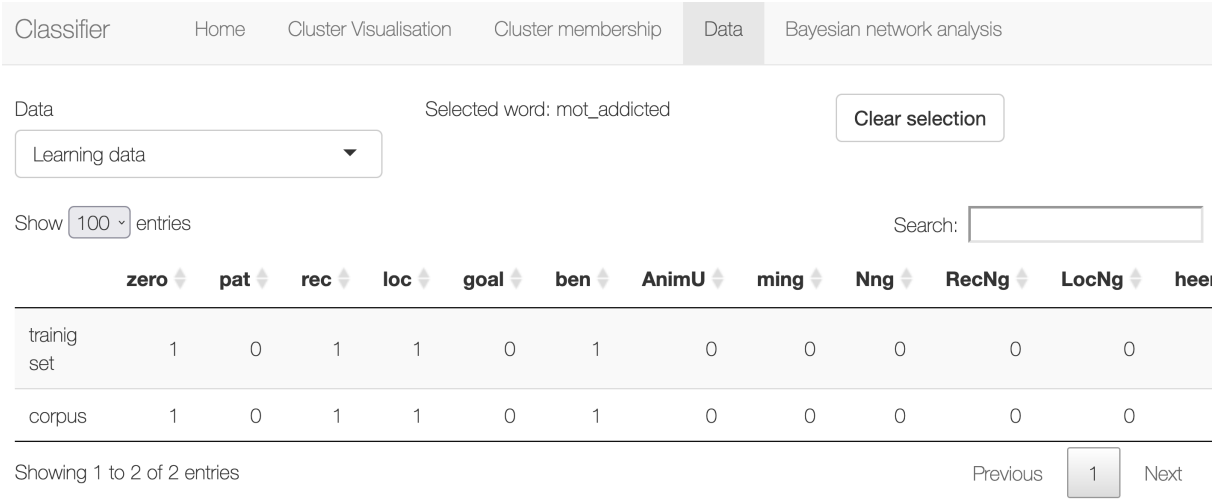

Figure 7: The data list comparison. This overview can be searched for individual verbs in either the *Learning set* or *Corpus data*, viewed in its entirety or navigated from the *Cluster Membership* tab.

aided by visualisation methods.

Because the corpus is not phonologically normalised, we are expanding the list of alternate forms. For example the verb *aquta* 'be blind' is also attested as *akuta* in the unglossed part of the corpus. In the first two versions there was also no explicit linking for mutating stems such as *meeng* 'wear (imperfective)' and as *meen* 'wear (perfective)'. The *Cluster list* allows us to discover further such candidates in the data and update the list of alternate forms.

3. Conclusion and future work

This paper presents an integrated workflow to support automatic verb classification in Abui, based on the morphological profile of the verb stem. The system is designed to support the advanced analysis of this complex grammatical feature of Abui that has been subject of a number of detailed investigations and continues to attract interest, especially in the context of the ongoing language shift and the growing influence of Alor Malay, which have been shown to lead to a gradual overhaul of the verb inflection system (Klamer and Saad, 2020; Saad, 2020b). In particular the *AVC-abz* system has got the following properties:

- Integrated native judgment and corpus data: Both types of data are handled as distinct types (*learning set* and *corpus data*. The learning set is manually curated with the assistance of native speakers and used as input for automated classifier.

- Corpus linguistics tool for extracting corpus information: The corpus is harvested for morphosyntactic information of verb stems following the methodology described in (Zamaraeva et al., 2017). The output is distributed into clusters.

- Mathematical tools for classification and feature relations: K-means and Naive Bayes are used to analyse the *learning set* and to assign the *corpus set* data to the clusters with the best fit.

- Version control: It is typical for documentation projects that the work is ongoing and the analysis is changing. The system is designed to work with different versions of the *learning set* and *corpus data* in order to compare the coverage of the corpus and the gradual improvement of the analysis.

- Interface supporting data interpretation by the expert: The interface will serve as an open-data platform to support future publications on the Abui verb class system. It enables the expert to investigate the detailed properties of the various verbs as well as the entire class system.

The system is also relevant to the *data coverage* question put forth by Nikolaus Himmelmann as: 'the aim of a language documentation is to provide a comprehensive record of the linguistic practices characteristic of a given speech community' (Himmelmann, 1998).

There are no standards to report corpus coverage cross-linguistically except simple metrics such as number of words or sentences, or to ascertain whether the aggregated corpus 'large enough' and presents a 'comprehensive record' of the language, speaking in Himmelmann's terms. While the question of 'large enough' may be a rhetorical distraction, it is practical to develop tools that can measure the corpus coverage of specific phenomena, whose aggregation can eventually answer the 'large enough' question. The AVC-abz addresses the question of corpus coverage locally; it measures the fit between the *learning set* and the *corpus data*. We can expect that at some point most verbs from the *corpus data* will always fit in the optimal number of clusters determined by the *learning set* whose size does not need to be increased anymore.

Our future work will focus on implementing a Bayesian Network to investigate the relationships between various inflectional categories. It hedges for the possibility that some features do correlate or are dependent. This information can be for example fed into the settings of the Naive Bayes classifier which can treat dependent or correlating features as one.

Another line of further improvement concerns the orthographic variation. The Abui corpus is partly sourced within the community and therefore contains multiple orthography standards and some dialectal variation. Abui speakers do not agree on a single orthography regarding velar and uvular stops ($k \sim q$ vs. k only), long vowels (e.g. $a \sim aa$ vs. a only), and tones (not written at all or written with accents). They also do not make a strict distinction between the single stem predicates and complex predicates containing light verbs, because in both cases the predicate (simple or complex) forms a single phonological word. The resulting variation in the unglossed part of the corpus is responsible for some noise in the classification, but the presented tool enables us to find such instances and to amend the orthographic profile of the given verb.

Acknowledgments

This article reports research carried out with the generous support of the Czech Science Foundation grant 20-18407S *Verb Class Analysis Accelerator for Low-Resource Languages - RoboCorp* (PI F Kratochvíl).

4. Bibliographical References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of COLING/ACL*, pages 86–90, Montreal. ACL.

Bender, E. M., Flickinger, D., and Oepen, S. (2002). The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. In John Carroll, et al., editors, *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan.

Bender, E. M., Drellishak, S., Fokkens, A., Poulson, L., and Saleem, S. (2010). Grammar Customization. *Research on Language & Computation*, pages 1–50. 10.1007/s11168-010-9070-1.

Bender, E. M., Crowgey, J., Goodman, M. W., and Xia, F. (2014). Learning grammar specifications from IGT: A case study of Chintang. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 43–53, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Fedden, S., Brown, D., Kratochvíl, F., Robinson, L. C., and Schapper, A. (2014). Variation in Pronominal Indexing: Lexical Stipulation vs. Referential Properties in Alor-Pantar Languages. *Studies in Language*, 38(1):44–79.

Fillmore, C. J. (1970). The grammar of hitting and breaking. In Roderick A. Jacobs et al., editors, *Readings in English Transformational Grammar*, pages 120–133. Waltham, MA: Ginn.

Georgi, R. (2016). *From Aari to Zulu: Massively Multilingual Creation of Language Tools using Interlin-*

ear Glossed Text. Ph.D. thesis, University of Washington.

Goodman, M. W. (2013). Generation of machine-readable morphological rules with human readable input. *UW Working Papers in Linguistics*, 30.

Iren Hartmann, et al., editors. (2013). *Valency Patterns Leipzig*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Himmelmann, N. P. (1998). Documentary and descriptive linguistics. *Linguistics*, 36(1):161–196.

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40.

Klamer, M. and Kratochvíl, F. (2018). The evolution of Differential Object Marking in Alor-Pantar languages. In Ilja Seržant et al., editors, *The Diachronic Typology of Differential Argument Marking*, Studies in Diversity Linguistics, pages 69–95. Language Science Press, Berlin.

Klamer, M. and Saad, G. (2020). Reduplication in Abui: A case of pattern extension. *Morphology*, 30:311–346.

Kratochvíl, F. and Delpada, B. (2015). Degrees of affectedness and verbal prefixation in Abui (Papuan). In Stefan Müller, editor, *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar, Nanyang Technological University (NTU), Singapore*, pages 216–233, Stanford, CA. CSLI Publications.

Kratochvíl, F., Moeljadi, D., Delpada, B., Kratochvíl, V., and Vomlel, J. (2021). Aspectual pairing and aspectual classes in Abui. *STUF - Language Typology and Universals*, 74(3-4):621–657.

Kratochvíl, F. (2007). *A grammar of Abui: a Papuan language of Alor*. LOT, Utrecht.

Kratochvíl, F. (2011). Transitivity in Abui. *Studies in Language*, 35(3):588–635.

Kratochvíl, F. (2014). Differential argument realization in Abui. *Linguistics*, 52(2):543–602.

Kratochvíl, F. (2022). Abui Corpus. Electronic Database: 162,000 words of natural speech, and 37,500 words of elicited material (February 2022). Palacký University Olomouc, Czech Republic.

Lepp, H., Zamaraeva, O., and Bender, E. M. (2019). Visualizing inferred morphotactic systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 127–131.

Levin, B. (2015). Semantics and pragmatics of argument alternations. *Annual Review of Linguistics*, 1(1):63–83.

Lewis, W. D. and Xia, F. (2008). Automatically identifying computationally relevant typological features. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 685–690, Hyderabad, India.

Merlo, P. and Stevenson, S. (2001). Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408.

O'Hara, K. (2008). A morphotactic infrastructure for a grammar customization system. Master's thesis, University of Washington.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Saad, G. (2020a). Abui. In Antoinette Schapper, editor, *The Papuan Languages of Timor, Alor and Pantar*, volume 3 of *Sketch Grammars*, chapter 5, pages 267–345. De Gruyter Mouton, Berlin.

Saad, G. M. (2020b). *Variation and change in Abui: The impact of Alor Malay on an indigenous language of Indonesia*. LOT, Amsterdam.

Siewierska, A. (2013). Verbal person marking. In Matthew S. Dryer et al., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

SIL International. (2015). Field Linguist's Toolbox. Lexicon and corpus management system with a parser and concordancer; URL: http://www-01.sil.org/computing/toolbox/documentation.htm.

SIL International. (2017). Sil Fieldworks. Lexicon and corpus management system with a parser and concordancer, URL: http://software.sil.org/fieldworks/.

Sun, L. and Korhonen, A. (2009). Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 638–647.

Wax, D. (2014). Automated grammar engineering for verbal morphology. Master's thesis, University of Washington.

Zamaraeva, O., Kratochvíl, F., Bender, E. M., Xia, F., and Howell, K. (2017). Computational Support for Finding Word Classes: A Case Study of Abui. In *Proceedings of ComputEL-2: 2nd Workshop on Computational Methods for Endangered Languages, Honolulu, Hawaii, March 6-7, 2017*. Association for Computational Linguistics (ACL).

Zamaraeva, O. (2016). Inferring Morphotactics from Interlinear Glossed Text: Combining Clustering and Precision Grammars. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 141–150, Berlin, Germany, August. Association for Computational Linguistics.