Decomposing Fusional Morphemes with Vector Embeddings

Michael Ginn and **Alexis Palmer**
University of Colorado
michael.ginn@colorado.edu and alexis.palmer@colorado.edu

Abstract

Distributional approaches have proven effective in modeling semantics and phonology through *vector embeddings*. We explore whether distributional representations can also effectively model morphological information. We train static vector embeddings over morphological sequences. Then, we explore morpheme categories for *fusional morphemes*, which encode multiple *linguistic dimensions*, and often have close relationships to other morphemes. We study whether the learned vector embeddings align with these linguistic dimensions, finding strong evidence that this is the case. Our work uses two low-resource languages, Uspanteko and Tsez, demonstrating that distributional morphological representations are effective even with limited data.

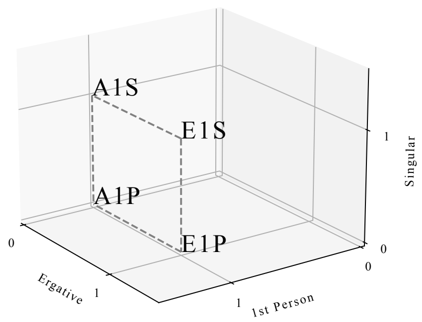

Figure 1: Morpheme glosses in a handcrafted linguistic feature space. Related glosses have predictable vector relationships. A=absolutive case, E=ergative case, P=plural number, S=singular number, 1=first person.

1 Introduction

Distributional semantics, which models the meanings of words according to the contexts in which they appear (Wittgenstein, 1953), has proven highly successful for language modeling. Generally, this has been achieved through **word embeddings**, which represent words with many-dimensional vectors (Turney and Pantel, 2010; Mikolov et al., 2013b; Levy and Goldberg, 2014b), and capture many linguistic patterns and regularities (Mikolov et al., 2013b; Levy and Goldberg, 2014a).

Linguistic research has suggested that this distributional approach can be effective across all units of language (Haas, 1954). Prior work (Silfverberg et al., 2018; Kolachina and Magyar, 2019) has explored a distributional approach to phonology, finding that embeddings for phonological units can capture predictable linguistic features and natural classes.

We explore whether this approach is also useful for morphology, hypothesizing that many grammatical morphemes can be described primarily by the contexts in which they appear. For example, a first person verbal affix might typically co-occur with first person pronouns, depending on the properties of the language being modeled.

We focus on groups of highly related morphemes, in particular instances of **fusional morphology**. Languages with fusional morphology include single morphemes that encode multiple grammatical features (as opposed to agglutinating morphology, where each morpheme corresponds to a single grammatical function). It is disputed whether languages exist with solely agglutinating or fusional morphological systems; rather, evidence suggests that many languages incorporate both processes (Plank, 1999; Haspelmath, 2009).

We compute morphological embeddings using standard vector embedding algorithms on morphological sequences from two low-resource languages, Uspanteko and Tsez (section 3). We compare these embeddings to handcrafted feature vectors based on the *linguistic dimensions* that make up the morphemes (see Figure 1). We find that there is a consistent correlation between the vector embedding space and this linguistic feature space.

Proceedings of the 21st SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 57–66
June 20, 2024 ©2024 Association for Computational Linguistics

2 Data and Languages

Data Format We utilize morpheme sequences from **interlinear glossed text** (IGT) data, a format commonly used in language documentation. An example of Uspanteko IGT is given in item 1.[1]

(1) Ti- j- ya' -tq -a' juntiir
INC- E3S- VT -PL -ENF ADV
They give us everything
(Pixabaj et al., 2007)

The first line records text in the target language. The second line, referred to as the *gloss line*, is a sequence of morphological glosses for each morpheme in the transcription, describing the morphological category and function of each morpheme. Often, stem morphemes may instead by glossed with a translation of the stem, however, in this work we use morphological category glosses as exemplified here (e.g. VT for the transitive verb stem *ya'*). The last line in an IGT example is generally a translation into English or a similarly high-resource language. We utilize only the gloss lines of IGT as morphological category sequences.

We use data from Ginn et al. (2023), which we have formatted in HuggingFace datasets, available online.[2] We use the train splits from Ginn et al. (2023), with 9,774 Uspanteko sentences and 7,116 Tsez sentences.

Languages Uspanteko (usp), or Uspantek, is an endangered Mayan language of Guatamala with around 6,000 speakers (Bennett et al., 2016). The language uses a system of absolutive and ergative affixes which generally attach to verbal stems (Coon, 2016). These affixes are fusional, encoding case (absolutive or ergative), number (singular or plural), and person (first, second, or third-person).

Tsez (ddo), or Dido, is a language in the Nakh-Daghestanian family, with around 14,000 speakers in Daghestan, Russia. Tsez utilizes a highly agglutinating and fusional morphological system, with morphemes often encoding two to five distinct linguistic dimensions. Our data is originally from the Tsez Annotated Corpus Project (Abdulaev et al., 2022; Abdulaev and Abdullaev, 2010).

3 Static Morphological Embeddings

We first investigate whether distributional representations are applicable to morphological sequences—that is, do the contexts that morphemes occur in reflect any meaningful linguistic relationships, and can we capture those relationships with distributional methods? To do this, we train embeddings over sequences of morphological categories from the gloss lines of the IGT from the corpora described in section 2.

We might also have trained embeddings over the morphemes themselves, rather than their glosses/categories. However, our corpora are rather small, and the majority of morphemes occur very rarely, making it difficult to induce meaningful representations. By studying sequences of morpheme categories, we can gain insight into broader morphological patterns, despite limited data.

3.1 Models

Following the approach used in Silfverberg et al. (2018), we consider two different models for learning morphological category embeddings. In all cases, directionality is not considered, so we treat neighboring glosses uniformly, regardless of whether they precede or follow the target gloss.

SVD We compute *positive pointwise mutual information* (PPMI) matrices for each morpheme category in some context window and calculate the *singular value decomposition* (SVD) (Bullinaria and Levy, 2007; Levy and Goldberg, 2014b). We truncate embeddings to some vector length d.

word2vec The word2vec (Mikolov et al., 2013a) model uses a shallow neural network, trained to predict the surrounding words in a sliding window, using the embedding layer as word representations. We use the `gensim` implementation[3] with the default parameters (including negative sampling) and experiment with both the skip-gram and continuous bag-of-words (CBOW) algorithms.

3.2 Experimental Settings

We train separate embedding models over the Uspanteko and Tsez morpheme sequences. For both model types (SVD and word2vec), we train models with vector sizes of 5 to 50 and window sizes of 1 to 10, for a total of 460 distinct runs for each language-model combination. We omit any glosses with fewer than five occurrances.

We believe it is important to report results across hyperparameter combinations, as this is an unsupervised task where it is difficult to tune hyperparameters, and using only a single combination of

[1] A full table of gloss definitions appears in Appendix B.

[2] https://huggingface.co/datasets/lecslab/usp-igt, https://huggingface.co/datasets/lecslab/ddo-igt

[3] https://radimrehurek.com/gensim/

Gloss	Most similar gloss		
	SVD	W2V (CBOW)	W2V (SG)
		Uspanteko	
A1P	A2P	A2S	A2S
E1P	A2P	E3	E3
S (noun)	AFI	SREL	SREL
VI	A2P	VT	VT
		Tsez	
DEM1.IPL	VOC	DEM2.IPL	DEM2.IPL
DEM2.IISG.OBL	VOC	DEM2.ISG.OBL	DEM2.ISG.OBL
POSS.ESS	COND.IRR	POSS.LAT	LAT
SUPER.ESS	IRR	IN.ESS	CONT.ESS

Table 1: For each gloss embedding, the gloss with the most similar embedding. Here we present a subset of interesting results, full results are in Appendix B.

parameters may produce results which are unrepresentative of the typical performance.

3.3 Results

3.3.1 Related glosses have similar embeddings

First, we investigate whether linguistically-related glosses tend to occur in similar contexts. For each gloss (e.g. A1S), and for every hyperparameter setting, we compute the most similar (distinct) embedding to the gloss's embedding, using cosine similarity. Then, for each gloss we select the most common similar gloss across hyperparameter settings. We highlight a subset of interesting results in Table 1, and report the full results in Appendix B.

We observe differences between the models. The `word2vec` models are far more likely to capture linguistically interesting similarities, while the SVD model does so much less reliably. In the `word2vec` results, closely related glosses, such as VI (intransitive verb stem) and VT (transitive verb stem) tend to be very similar. Both `word2vec` models predict SREL (relational noun) as the most similar gloss to S (noun). Additionally, fusional morpheme glosses such as E1S (ergative first-person singular) tend to be similar to other fusional glosses with the same features, such as E2S (ergative second-person singular). The results for Tsez show similar patterning, with `word2vec` models more closely aligning glosses representing related categories.

3.3.2 Gloss embedding spaces correlate with linguistic feature spaces

Following Silfverberg et al. (2018), we conduct a quantitative measurement in order to understand whether the geometry of the embedding space correlates with a space defined by manually chosen linguistic features. We do not make any assumptions about the magnitude or orientation of embedding vectors; rather, we focus on the cosine similarity scores between embedding vector pairs.

Specifically, we assign vectors to the fusional morphemes in each dataset, using the linguistic dimensions defined in the UniMorph schema (Kirov et al., 2016) as features. Unlike the phonological feature spaces of Silfverberg et al. (2018), it is difficult to decompose all glosses into a single set of linguistic dimensions, as many glosses are completely unrelated. Instead, we focus on the subset of morpheme glosses which share clear features. Each linguistic feature value (e.g. ergative case) is represented as a binary dimension, as in Figure 2. We describe the glosses and linguistic dimensions in detail in Appendix B.

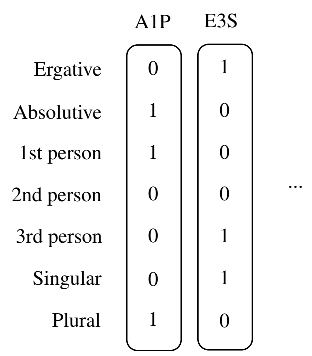

Figure 2: Each morpheme gloss is assigned a handcrafted linguistic feature vector, based on linguistic dimensions from the Unimorph schema. Two examples in Uspanteko are shown here.

For a pair of fusional morpheme glosses, we compute the cosine similarity of the linguistic feature vectors for each gloss. We also compute the cosine similarity for the same glosses using the embedding vectors from the embedding model. We aggregate these similarity measurements across all pairs of glosses that have at least one feature in common. Glosses without any features in common are orthogonal in the linguistic space, hence similarity will be 0. As embedding vectors will generally never have a similarity of 0, we found this added significant noise to the correlation calculation.

Then, we compute the linear correlation coefficient between the linguistic space similarities and the embedding space similarities. As a baseline, we select a random vector in the embedding space for each gloss vector, compute similarities, and calculate the correlation coefficient with the linguistic space similarities. We conduct this process over the hyperparameter combinations described above and report summary results in Table 2 and box plots in Figure 3 and Figure 4.

	Mean / max correlation coefficient r		
	SVD	W2V (CBOW)	W2V (SG)
Uspanteko			
Random	0.05 / 0.49	-0.06 / 0.27	-0.03 / 0.35
True	0.26 / **0.68**	0.19 / 0.42	**0.36** / 0.50
Tsez			
Random	0.02 / 0.10	-0.04 / 0.08	-0.04 / 0.06
True	**0.21** / **0.27**	0.08 / 0.13	0.12 / 0.19

Table 2: Mean / max correlations between linguistic feature space and embedding feature spaces, across hyperparameters.

Findings Broadly, we find that the correlations between the linguistic feature spaces and the vector embedding spaces are greater than the correlations with randomly-selected vector embedding spaces, with the SVD models achieving the highest max correlation across languages. We conduct a paired t-test between the random and true correlation values for each model and language, and find that there is a statistically significant difference in every case with $p < 0.001$

The mean correlations are still fairly low—this is likely due in part to the small size of the dataset, but may also indicate that the models are learning relationships between morphemes other than the linguistic dimensions we specify. Future work could investigate these vector spaces more thoroughly to search for novel morphological relationships.

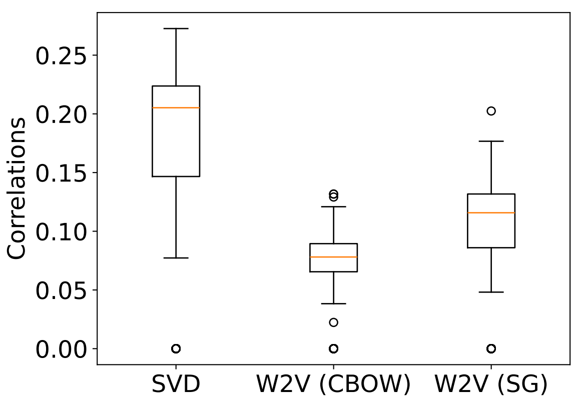

Figure 3: Box plots for Tsez correlation values across hyperparameter values.

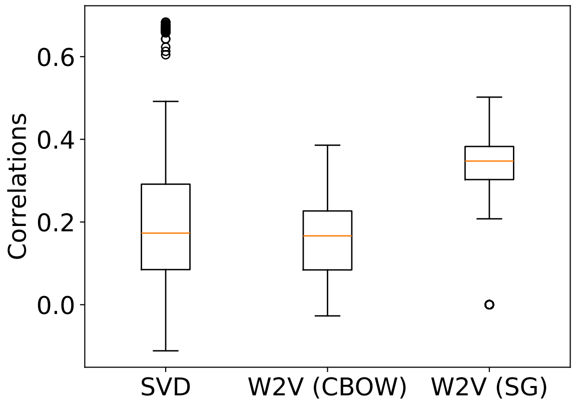

Figure 4: Box plots for Uspanteko correlation values across hyperparameter values.

Hyperparameters Not all hyperparameter values perform equally well. We report heatmaps for each model across window size and vector size in Figure 5 and Appendix A. For SVD models, correlation with the linguistic space is maximized with small window sizes (1-2) and decreases significantly with greater window sizes, indicating that the features captured by our linguistic dimensions are generally locally predictable. On the other hand, the `word2vec` models seem to have more consistent performance across window sizes, perhaps indicating that the models are more robust against the noise induced with larger windows. None of the models show significant differences across vector sizes, although the SVD models perform poorly with large windows and very small vector sizes.

4 Related Work

Word embeddings (Turney and Pantel, 2010; Mikolov et al., 2013a,b; Levy and Goldberg, 2014b) have been widely successful in NLP, capturing semantic relationships in many-dimensional vector representations.

Vector embeddings have been applied to phonology, where *phone embeddings* have been used to capture phonetic relationships (Silfverberg et al., 2018; Kolachina and Magyar, 2019; Mayer, 2020).

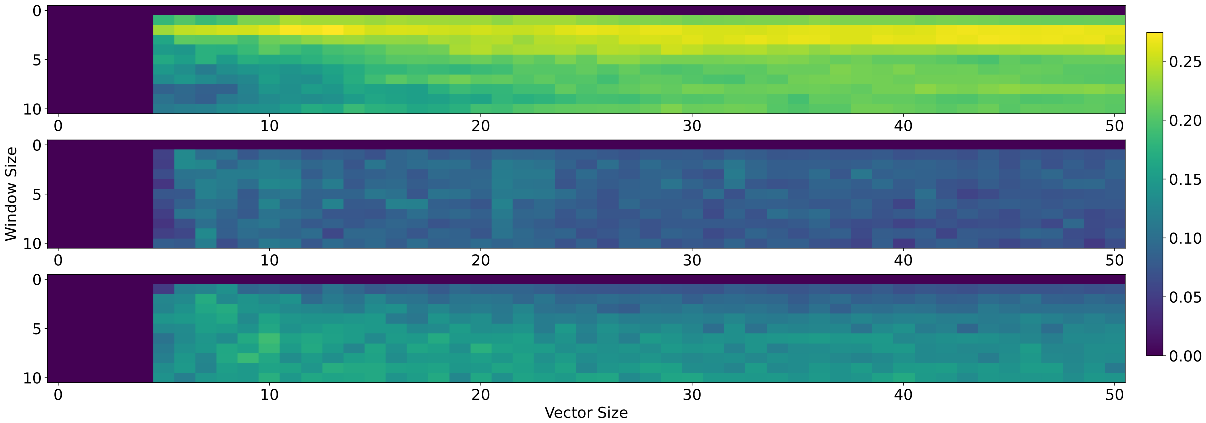

Figure 5: Heatmaps for Tsez of vector space correlation over hyperparameters between the linguistic feature space and the embedding spaces produced by the SVD (top), CBOW (middle), and Skip-gram (bottom) models.

Morphological information has been integrated into word embeddings to improve representations in morphologically-rich languages (Cao and Rei 2016; Edmiston and Stratos 2018; Ataman and Federico 2018; Schwartz et al. 2022, inter alia). To our knowledge, this is the first work that explores a distinct level of morpheme embeddings.

5 Conclusion

We find evidence that distributional vector representations of morpheme categories capture linguistic regularities, even with limited data. Broadly, morphological features such as number, case, and person seem to correlate with the contexts those morphemes appear in. We suggest that distributional morpheme representations are a viable model for morphology, particularly in languages with highly-productive, fusional morphemes.

This research is motivated primarily by linguistic understanding; that is, we are interested in determining whether morpheme contexts have predictable relationships. However, we suggest these findings could be applied in future research to more practical ends. For example, a linguist might use this approach to investigate a hypothesis about the relatedness of certain morphemes, providing for data-driven, large-scale evidence. Alternately, an NLP practioner could use these findings in a task such as morpheme glossing (Ginn et al., 2023) to design models that utilize shared features to make predictions.

6 Limitations

Our research utilizes morphological datasets from two distinct languages. However, considering the linguistic diversity of the world's languages, we expect results may vary across additional languages. In particular, languages without fusional morphology may not show strong linguistic correlations, like we observed in this work.

Acknowledgments

This work utilized the Blanca condo computing resource at the University of Colorado Boulder. Blanca is jointly funded by computing users and the University of Colorado Boulder. Portions of this work were supported by the National Science Foundation under Grant No. 2149404, "CAREER: From One Language to Another". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

A. K. Abdulaev and I. K. Abdullaev, editors. 2010. *Cezyas folklor/Dido (Tsez) folklore/Didojskij (cezskij) fol ́klor*. "Lotos", Leipzig–Makhachkala.

A.K. Abdulaev, I.K. Abdullaev, André Müller, Evgeniya Zhivotova, and Bernard Comrie. 2022. The Tsez Annotated Corpus Project.

Duygu Ataman and Marcello Federico. 2018. Compositional Representation of Morphologically-Rich Input for Neural Machine Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 305–311, Melbourne, Australia. Association for Computational Linguistics.

Ryan Bennett, Jessica Coon, and Robert Henderson. 2016. Introduction to Mayan Linguistics. *Lang. Linguistics Compass*, 10:455–468.

John A Bullinaria and Joseph P Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39:510–526.

Kris Cao and Marek Rei. 2016. A joint model for word embedding and word morphology. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 18–26, Berlin, Germany. Association for Computational Linguistics.

Jessica Coon. 2016. Mayan Morphosyntax: Mayan Morphosyntax. *Language and Linguistics Compass*, 10(10):515–550.

Daniel Edmiston and Karl Stratos. 2018. Compositional morpheme embeddings with affixes as functions and stems as arguments. In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 1–5, Melbourne, Australia. Association for Computational Linguistics.

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201, Toronto, Canada. Association for Computational Linguistics.

W. Haas. 1954. On defining linguistic units. *Transactions of the Philological Society*, 53(1):54–84.

Martin Haspelmath. 2009. *An Empirical Test of the Agglutination Hypothesis*, pages 13–29. Springer Netherlands, Dordrecht.

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of Wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3121–3126, Portorož, Slovenia. European Language Resources Association (ELRA).

Sudheer Kolachina and Lilla Magyar. 2019. What do phone embeddings learn about phonology? In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 160–169.

Omer Levy and Yoav Goldberg. 2014a. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Connor Mayer. 2020. An algorithm for learning phonological classes from distributional similarity. *Phonology*, 37(1):91–131.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Telma Can Pixabaj, Miguel Angel Vicente Méndez, Marıa Vicente Méndez, and Oswaldo Ajcot Damián. 2007. Text collections in four mayan languages. *Archived in The Archive of the Indigenous Languages of Latin America*.

Frans Plank. 1999. Split morphology: How agglutination and flexion mix. *Linguistic Typology*, 3:279–340.

Lane Schwartz, Coleman Haley, and Francis Tyers. 2022. How to encode arbitrarily complex morphology in word embeddings, no corpus needed. In *Proceedings of the first workshop on NLP applications to field linguistics*, pages 64–76, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Miikka Silfverberg, Lingshuang Jack Mao, and Mans Hulden. 2018. Sound analogies with phoneme embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 136–144.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Blackwell, Oxford.

A Uspanteko Heatmap

We provide the correlation heatmap for Uspanteko, similar to the Tsez figure provided in the main paper in Figure 6.

B Glosses

We report a complete list of the glosses in each language in Table 3 and Table 4.

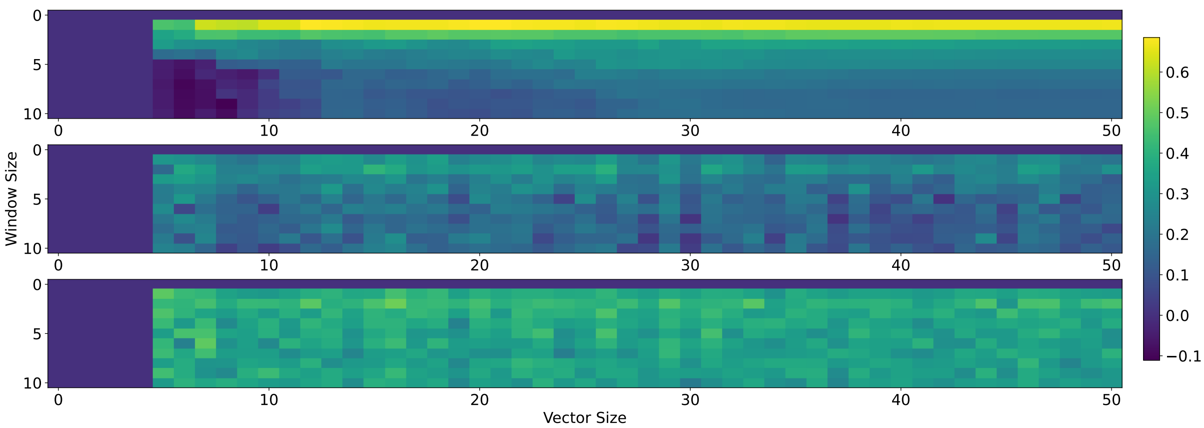

Figure 6: Heatmaps for Uspanteko of vector space correlation over hyperparameters between the linguistic feature space and the embedding spaces produced by the SVD (top), CBOW (middle), and Skip-gram (bottom) models.

Gloss	Label	Count	Features	Most similar gloss SVD	CBOW	SG
A1P	Absolutive 1P Pl	110	Abs., 1st, Pl.	A2P	A2S	A2S
A1S	Absolutive 1P Sing	347	Abs., 1st, Sing.	REC	A2S	A2S
A2S	Absolutive 2P Sing	127	Abs., 2nd, Sing.	DIM	A1P	A1P
ADJ	Adjective Stem	1017		APLI	NUM	ITS
ADV	Adverb Stem	5830		APLI	PART	PART
AFE	Affective	116		A2P	PREP	PREP
AFI	Positive	208		E3P	PART	DEM
AGT	Agentive	100		A2P	E2	E2
AP	Antipassive	339		A2P	E2S	E2S
ART	Article	973		INT	NUM	NUM
CAU	Causative	19		PRG	GNT	RFX
CLAS	Classifier	155		REC	NOM	NOM
COM	Completive	2304		NOM	INC	PP
COND	Conditional	59		REC	IMP	PRG
CONJ	Conjunction	1152		A2P	VOC	VOC
DEM	Dem.	2116		APLI	AFI	AFI
DIM	Diminutive	797		A2S	ART	NUM
DIR	Directional	687		A2P	PAS	PAS
E1P	Ergative 1P Pl	1370	Erg., 1st, Pl.	A2P	E3	E3S
E1S	Ergative 1P Sing	709	Erg., 1st, Sing.	NOM	E2S	E2S
E2	Ergative 2P	16	Erg., 2nd	NOM	INS	RFX
E2P	Ergative 2P Pl	16	Erg., 2nd, Pl.	ART	INS	E2
E2S	Ergative 2P Sing	564	Erg., 2nd, Sing.	A2P	AP	AP
E3	Ergative 3P	385	Erg., 3rd	A2P	E3S	E3S
E3P	Ergative 3P Pl	32	Erg., 3rd, Pl.	NOM	E2P	E2P
E3S	Ergative 3P Sing	3118	Erg., 3rd, Sing.	NOM	E3	E3
ENF	Emphasis	1464		A2P	A1P	IMP
EXS	Existential	661		A1P	NUM	NUM
GNT	Demonym	20		TRN	INS	RFX
IMP	Imperative	67		EXS	COND	COND
INC	Incompletive	2742		NOM	COM	SC
INS	Instrumental	37		A2P	GNT	E2P
INT	Interrogative	343		ART	NEG	NEG
ITR	Intransitive	73		A2P	E2	RFX
ITS	Intensifier	244		GNT	AFI	ADJ
MED	Measure	66		A2P	POS	AGT
MOV	Auxiliary	141		REC	AGT	AGT
NEG	Negative	1130		REC	INT	INT
NOM	Proper Name	167		PAS	CLAS	CLAS
NUM	Numeral	1029		APLI	ART	MED
PART	Particle	3153		A2P	ADV	ADV
PAS	Passive	276		A2P	DIR	E3P
PL	Pl	2094		DEM	PREP	PREP
POS	Positional	83		E2P	MED	GNT
PP	Perfect Participle	127		REC	AGT	AGT
PREP	Preposition Stem	1605		A2P	PL	AFE
PRG	Progressive	42		CAU	GNT	TRN
PRON	Pronoun	1674		REC	INT	A2S
RFX	Reflexive	8		INT	GNT	TRN
S	Noun Stem	6962		AFI	SREL	SREL
SAB	Abstract Noun Stem	158		CONJ	MED	INS
SC	Category Suffix	1018		A2P	ENF	SV
SREL	Relative Noun	1890		TRN	S	S
SV	Verbal Noun Stem	88		E1	INS	INS
TAM	Tense-Aspect-Mood	128		APLI	SV	SV
TOP	Proper Noun Stem	108		A2P	MED	GNT
TRN	Applicative	7		TOP	GNT	RFX
VI	Intransitive Verb Stem	3125		A2P	VT	VT
VOC	Vocative	750		A2P	CONJ	CONJ
VT	Transitive Verb Stem	5024		NOM	VI	VI

Table 3: All of the glosses in Uspanteko, along with a description, the total number of occurrances, and a list of positive features in the linguistic vector representations.

Gloss	Label	Most similar gloss SVD	CBOW	SG
AD.ABL	Position At, Ablative	PST.UNW	APUD.ABL	AD.VERS
AD.ESS	Position At, Essive	APUD.VERS.DIST	SUB.ABL	SUB.ABL
AD.LAT	Position At, Lative	COND.IRR	IN.ESS	IN.ESS
AD.VERS	Position At, Versative	POSS.ESS.DIST	SUPER.VERS	CONT.VERS
AD.VERS.DIST	Position At, Versative, Distal	PROHIB	APUD.ABL	CONT.ABL.DIST
ANT.CVB	Anterior, Converb	COND.IRR	IMM.ANT.CVB	IMM.ANT.CVB
APUD.ABL	Pos. Near, Ablative	DEM2.IIPL	INT	AD.VERS.DIST
APUD.ESS	Pos. Near, Essive	PST.UNW	APUD.VERS	APUD.LAT
APUD.LAT	Pos. Near, Lative	APUD.ABL.DIST	APUD.VERS	APUD.VERS
APUD.VERS	Pos. Near, Versative	PST.UNW	APUD.LAT	APUD.LAT
APUD.VERS.DIST	Pos. Near, Versative, Distal	SUPER.LAT.DIST	DEM3.SG	LOC.ORIG
ATTR	Attributive	NEG.PRSPRT.OBL	ATTR.OBL	RES.PRT.OBL
ATTR.OBL	Attributive, Oblique	SUPER.ESS.DIST	ATTR	GEN2
CNC.CVB	Concessive, Converb	DEM2.IIPL	COND	COND
CND	Conditional	SUPER.LAT.DIST	CONT.ABL.DIST	IN.LAT.DIST
CND.CVB	Conditional, Converb	DIST	PRS.PRT	COND
CND.CVB.IRR	Conditional, Converb, Irrealis	SUPER.LAT.DIST	COND	DEM3.SG
COND	Conditional	SUPER.LAT.DIST	DEM3.SG	NEG.PRS.PRT
COND.IRR	Conditional, Irrealis	SUPER.LAT.DIST	DUB	INDEF
CONT.ABL	Pos. Among, Ablative	GER.PURP	IN.ESS	GEN1
CONT.ABL.DIST	Pos. Among, Ablative, Distal	EQU1	APUD.ABL	IN.LAT.DIST
CONT.ESS	Pos. Among, Essive	SUPER.ESS.DIST	GEN1	POSS.ABL
CONT.LAT	Pos. Among, Lative	NEG.PRS.PRT.OBL	SUB.ESS	SUB.ESS
CONT.VERS	Pos. Among, Versative	NEG.PRSPRT	AD.VERS	IN.ABL
CONT.VERS.DIST	Pos. Among, Versative, Distal	SUPER.LAT.DIST	IN.ESS.DIST	SUPER.ABL.DIST
CSL.CVB	Causal, Converb	IN.VERS.DIST	INF	NEG.PST.UNW
DEF	Definite	SUPER.ESS.DIST	AD.LAT	SUB.LAT
DEM1.IIPL	C1 Dem. 2nd N Pl	PST.UNW	POSS.VERS	DEM1.IIPL.OBL
DEM1.IIPL.OBL	C1 Dem. 2nd N Pl, Oblique	VOC	DEM2.IPL.OBL	DEM1.IIPL
DEM1.IISG.OBL	C1 Dem. 2nd N Sing, Oblique	SUPER.LAT.DIST	DEM2.ISG.OBL	DEM3.IISG.OBL
DEM1.IPL	C1 Dem. 1st N Pl	VOC	DEM2.IPL	DEM2.IPL
DEM1.IPL.OBL	C1 Dem. 1st N Pl, Oblique	RES.PRT.OBL	DEM2.IPL.OBL	DEM1.IPL
DEM1.ISG.OBL	C1 Dem. 1st N Sing, Oblique	NEG.PST.UNW	DEM2.ISG.OBL	DEM2.IISG.OBL
DEM1.SG	C1 Dem. Sing	APUD.VERS.DIST	II	DEM4.SG
DEM2.IIPL.OBL	C2 Dem. 2nd N Pl, Oblique	DEM1.IISG	DEM3.IISG.OBL	DEM3.IISG.OBL
DEM2.IISG	C2 Dem. 2nd N Sing	APUD.ABL.DIST	PROHIB	DEM1.SG
DEM2.IISG.OBL	C2 Dem. 2nd N Sing, Oblique	VOC	DEM2.ISG.OBL	DEM2.ISG.OBL
DEM2.IPL	C2 Dem. 1st N Pl	CND.CVB	DEM1.IPL	DEM1.IPL
DEM2.IPL.OBL	C2 Dem. 1st N Pl, Oblique	NEG.PRS.PRT	DEM1.IIPL.OBL	DEM2.IPL
DEM2.ISG	C2 Dem. 1st N Sing	LNK	IN.LAT	IN.ESS.DIST
DEM2.ISG.OBL	C2 Dem. 1st N Sing, Oblique	NEG.PST.UNW	DEM1.ISG.OBL	DEM2.IISG.OBL
DEM2.PL	C2 Dem. 2nd N Pl	DEM3.IPL	POSS.VERS	CONT.VERS.DIST
DEM3.IISG.OBL	C3 Dem. 2nd N Sing, Oblique	SUPER.LAT.DIST	DEM2.IIPL.OBL	INTS
DEM3.SG	C3 Dem. Sing	SUPER.LAT.DIST	COND	COND.IRR
DEM4.IISG.OBL	C4 Dem. 2nd N Sing, Oblique	SUPER.LAT.DIST	DEM1.IIPL.OBL	LCV
DEM4.ISG.OBL	C4 Dem. 1st N Sing, Oblique	NEG.PST.UNW	DEM3.IISG.OBL	CONT.ABL.DIST
DEM4.SG	C4 Dem., Sing	SUPER.LAT.DIST	DEM4.ISG.OBL	DEM4.ISG.OBL
FUT.CVB	Future, Converb	SUB.ESS.DIST	NEG.FUT.DEF	NEG.PRS.PRT
FUT.DEF	Future, Definite	LNK	NEG.FUT	NEG.FUT
I.PL	1st Noun, Plural	CND.CVB	DEM2.IPL	DEM2.IPL
II	2nd Noun	LNK	DEM1.SG	DEM2.IISG
II.PL	2nd Noun, Plural	CND.CVB	IV.PL	DEM1.IIPL
III	3rd Noun	LNK	DEM2.ISG	DEM1.SG
III.PL	3rd Noun, Plural	SUB.ESS.DIST	II.PL	DEM1.IIPL
IMM.ANT.CVB	Immediate, Anterior, Converb	NEG.PST.UNW	POST.CVB	POST.CVB
IN.ABL	Position In, Ablative	NEG.PST.UNW	IN.ALL	CONT.VERS
IN.ABL.DIST	Position In, Ablative, Distal	CND.CVB	IN.ESS.DIST	CONT.VERS
IN.ALL	Position In, Allative	CND.CVB	IN.LAT	IN.VERS.DIST
IN.ESS	Position In, Essive	POSS.ESS.DIST	AD.LAT	AD.LAT
IN.ESS.DIST	Position In, Essive, Distal	POSS.ESS.DIST	APUD.ABL	CONT.ABL.DIST

Table 4: All of the glosses in Tsez, along with a description, the total number of occurrances, and a list of positive features in the linguistic vector representations. C# Dem.=class of demonstratives. The I, II, etc., morphemes indicate the four noun classes of Tsez.

Gloss	Label	Most similar gloss SVD	CBOW	SG
IN.LAT	Position In, Lative	CND.CVB	IN.ALL	IN.ABL.DIST
IN.LAT.DIST	Position In, Lative, Distal	SUPER.LAT.DIST	IN.ESS.DIST	CND
IN.VERS	Position In, Versative	OSS.ESS.DIST	CONT.LAT	AD.VERS
IN.VERS.DIST	Position In, Versative, Distal	SEQ	IN.ESS.DIST	IN.ESS.DIST
INT	Interrogative	APUD.VERS.DIST	APUD.ABL	IN.LAT.DIST
IPFV.CVB	Imperfective, Converb	POSS.ESS.DIST	TERM	TERM
IV	4th Noun	SUPER.LAT.DIST	III	NMLZ
IV.PL	4th Noun, Plural	POSS.ABL.DIST	II.PL	II.PL
LAT	Lative	PST.UNW	POSS.ESS	POSS.ESS
LCV	Locative	GER.PURP	LCV.CVB	POSS.VERS
LCV.CVB	Locative, Converb	PFV.CVB.INT	LCV	LCV
LOC.ORIG	Locative, Origin	GER.PURP	CONT.VERS.DIST	POSS.ABL.DIST
NEG	Negative	SUB.ESS.DIST	Q	Q
NEG.FUT	Negative, Future	SUB.VERS	FUT.DEF	FUT.DEF
NEG.FUT.CVB	Negative, Future, Converb	PST.UNW	COND.IRR	NEG.FUT
NEG.FUT.DEF	Negative, Future, Definite	SUPER.LAT.DIST	PST.WIT.Q	NEG.PRS.PRT
NEG.PRS.PRT	Negative, Present Participle	SUPER.LAT.DIST	NEG.PST.UNW	NEG.PST.UNW
NEG.PRS.PRT.OBL	Neg., Pres. Part., Oblique	APUD.ABL.DIST	CONT.VERS.DIST	POSS.ABL.DIST
NEG.PST.CVB	Negative, Past, Converb	SUB.ESS.DIST	TERM	NEG.PST.UNW
NEG.PST.UNW	Neg., Past, Unwitnessed	DEM2.ISG.OBL	POT	NEG.PRS.PRT
NEG.PST.WIT	Neg., Past, Witnessed	NEG.PST.UNW	Q	PST.WIT.INT
PCT.CVB	Perfective, Converb	IN.VERS	DEM3.SG	POSS.ABL.DIST
PFV.CVB	Perfective, Converb	VOC	EMPH	IN.VERS.DIST
PL	Plural	SUB.ESS.DIST	DEM1.IPL	DEM1.IIPL
POSS.ABL	Position Vertical, Ablative	PST.UNW	APUD.LAT	APUD.LAT
POSS.ABL.DIST	Pos. Vert., Ablative, Distal	SUPER.LAT.DIST	INTS	LOC.ORIG
POSS.ESS	Position Vertical, Essive	APUD.ABL.DIST	POSS.LAT	LAT
POSS.LAT	Position Vertical, Lative	POSS.ESS.DIST	POSS.ESS	GEN1
POSS.VERS	Position Vertical, Versative	COND.IRR	DEM2.PL	AD.VERS.DIST
POST.CVB	Posterior, Converb	LNK	IMM.ANT.CVB	IMM.ANT.CVB
PRS	Present	SUPER.LAT.DIST	FUT.DEF	PST.WIT.Q
PRS.PRT	Present Participle	NEG.PST.UNW	NEG.FUT	NEG.FUT
PRS.PRT.OBL	Present Participle, Oblique	POSS.ESS.DIST	DEM2.IIPL.OBL	DEM4.ISG.OBL
PST.PRT	Past, Participle	DEF1.IISG	ATTR	ATTR
PST.UNW	Past, Unwitnessed	POSS.ESS.DIST	ANT.CVB	ANT.CVB
PST.WIT	Past, Witnessed	PST.UNW	IMPR	NEG.PST.WIT
PST.WIT.INT	Past, Witnessed, Interr.	NEG.PST.UNW	NEG.PST.WIT	NEG.PST.WIT
PST.WIT.Q	Past, Witnessed, Question	IRR	NEG.FUT.DEF	DEM3.IISG.OBL
PURP.CVB	Purposive, Converb	ATTR.OBL	COND	PCT.CVB
Q	Question	AD.ABL.DIST	NEG.PST.WIT	NEG.PST.WIT
RES.PRT	Resultative Participle	SUPER.LAT.DIST	INF	PST.WIT.Q
RES.PRT.OBL	Res. Part., Oblique	LHUN	DEM3.SG	POSS.ABL.DIST
SIM.CVB	Simultaneous Converb	CND.CVB	IMM.ANT.CVB	ANT.CVB
SUB.ABL	Position Under, Ablative	POSS.ESS.DIST	AD.ESS	AD.ESS
SUB.ESS	Position Under, Essive	GER.PURP	CONT.LAT	APUD.ABL
SUB.LAT	Position Under, Lative	SUPER.ESS.DIST	APUD.ABL	CONT.ABL.DIST
SUPER.ABL	Position Under, Ablative	IN.LAT.DIST	CONT.ESS	CONT.ESS
SUPER.ABL.DIST	Pos. Under, Ablative, Distal	NEG.PRSPRT.OBL	INT	POSS.ABL.DIST
SUPER.ESS	Position Above, Essive	IRR	IN.ESS	CONT.ESS
SUPER.LAT	Position Above, Lative	POSS.ESS.DIST	IN.ABL	LCV.CVB
SUPER.VERS	Position Above, Versative	IRR	AD.VERS	IN.ESS.DIST
SUPER.VERS.DIST	Pos. Above, Versative, Distal	GER.PURP	APUD.ABL	APUD.VERS.DIST

Table 5: Tsez glosses (cont.)