A Classifier of Word-Level Variants in Witnesses of Biblical Hebrew Manuscripts

Iglika Nikolova-Stoupak[†,‡] **Maxime Amblard**[†]
Sophie Robert-Hayek[‡] **Frédérique Rey**[‡]
[†]LORIA, UMR 7503, Université de Lorraine, CNRS, Inria, 54000 Nancy, France
[‡]Research Centre Écritures, EA 3943, Université de Lorraine, 57000 Metz, France
{firstname.surname}@univ-lorraine.fr

Abstract

The current project is inscribed within the field of stemmatology or the study and/or reconstruction of textual transmission based on the relationship between the available witnesses of given texts. In particular, the variants (differences) at the word-level in manuscripts written in Biblical Hebrew are addressed. A dataset based on the Book of Ben Sira is manually annotated for the following variant categories: 'plus/minus', 'inversion', 'morphological', 'lexical' or 'unclassifiable'. A strong classifier (F1 value of 0.80) is then trained to predict these categories in collated (aligned) pairs of witnesses. The classifier is non-neural and makes use of the two words themselves as well as part-of-speech (POS) tags, hand-crafted rules per category, and additional synthetically derived data. Other models experimented with include neural ones based on the state-of-the-art model for Modern Hebrew, DictaBERT. Other features whose relevance is tested are different types of morphological information pertaining to the word pairs and the Levenshtein distance between the words within a pair. The strongest classifier as well as the used data are made publicly available. Coincidentally, the corelation between two sets of morphological labels is investigated: professionally established as per the Qumran-Digital online library and automatically derived with the sub-model DictaBERT-morph.

1 Introduction

Stemmatology, situated within the field of textual criticism, studies the genealogy of texts (Roelli, 2020). Within its framework, textual witnesses (i.e. extant versions of the same text) are aligned in a process known as 'collation' and compared to one another. In particular, it is assumed that variant differences (sometimes referred to as 'errors') associated with discrete witnesses give out important information about their relationship. If the same error is shared by two witnesses, and it is unlikely to have been made independently by the two scribes, then one of the witnesses is assumed to have been derived from the other. Stemmatology traditionally concerns academic disciplines such as classical philology and Biblical studies. It is associated with a number of 'schools', notably the 'German' one (represented by Karl Lachmann), which focuses mainly on intertextual connections, and the 'French' one (represented by Joseph Bédier), which also accords importance to a text's historical and cultural framework. More recently, the so-called 'new philology' proposes to move from the genealogical model to a study of each textual witness and its specific context (Cerquiglini, 1983; Jansen, 1990).

Multidisciplinarity is crucial to the practice and reliability of stemmatology, especially in the current digital era. Computing solutions have been used within the field since as early as the 1950s, due to its clear algorithmic nature (Heikkilä, 2023). Indeed, automation can be successfully applied to a number of aspects of the discipline, such as collation, statistics related to textual variants and even the ultimate construction of genealogical trees of texts. However, expert knowledge pertaining to the concerned academic disciplines and optimal communication within collaborating teams are crucial. This project is produced by a team specialising in diverse fields such as natural language processing (NLP), stemmatology, theology and Hebrew studies. The associated work seeks to establish exemplar practices in the application of contemporary NLP techniques to the classification of ancient manuscripts. Specifically, texts in Biblical Hebrew (and in particular, the Dead Sea Scrolls, whose age is estimated as 3rd century BCE - 1st century CE) are approached. Following elaborate manual annotation, classifiers of the variants between textual witnesses are trained. The strongest classifier is a non-neural (Random Forests) one that utilises the annotated data as well as part-of-speech (POS) tags, a limited amount of synthetic data and several hand-

Findings of the Association for Computational Linguistics: ACL 2025, pages 21313–21329

crafted rules that increase the probability of specific categories being predicted.

The long-term objective of our project is to establish a system that helps reconstruct the genealogical link between discrete manuscripts. An important step therein is to fully consider the discrepancies between them. At an atomic level, the differences between word pairs (omission/addition of a letter, replacement with a synonym, etc.) need to be not only counted but also categorised, as they may imply different levels and types of inter-textual relations. The present work focuses on this initial step, providing a classifier that achieves an F1 score of 0.80 while making use of original and synthetic data and taking into consideration the specificity of the Hebrew language.

The dataset of professionally annotated variants, the derived synthetic datasets and the strongest achieved classifier model are made available at: https://gitlab.inria.fr/semagramme/sherbet/

2 Background

The following discussion will consider the linguistic features of the Hebrew language as well as existing relevant NLP tools.

2.1 Varieties of Hebrew

Hebrew is a Northwest Semitic language which is read from left to right and makes use of an *abjad* writing system; that is to say, only consonants are typically represented. Diacritical signs (*nikkud*) may be added in order to denote vowel sounds and thus facilitate reading. The language is commonly described as morpho-synctactic (Khan et al., 2013). The general meaning of a word is carried by its typically three-letter root. Prepositions and conjunctions are prefixed and possessive pronouns are suffixed to the word they modify.

A common dichotomy exists between Modern and Classical (Ancient) Hebrew i.e. the language spoken in Israel today versus the language of the Hebrew Bible. Whilst morphology is the least altered aspect of the language (Taylor, 2019), its lexicon has been significantly enriched so as to include modern terms and concepts. Modern Hebrew makes use of a number of words whose roots can be traced to Biblical Hebrew but whose meaning has been adapted. For instance, the word יָטוּשׁ is a hapax legomenon found in the Book of Job, which describes the flight of an eagle; today, it means 'to fly on an airplane' (Pritz, 2016). Other notable linguistic developments include the loss or decline of some verb forms and tenses, such as the 'consecutive tenses'[1], the lengthened imperative and the jussive; the substitution of the conjunction אֲשֶׁר with שֶׁ; and the no longer compulsory question particle הַ (Khan et al., 2013). The majority of these developments have in fact been gradual and are traceable throughout the multiple defined sub-periods associated with the language, such as Archaic Hebrew, Classical Hebrew, Late Biblical Hebrew, Rabbinical Hebrew and Medieval Hebrew (Khan et al., 2013; Pérez Fernández and Elwolde, 1999; Schniedewind, 2013). A major development that can be traced to a specific historical point is the inclusion of diacritical signs in the writing system in the Masoretic era (7th-10th centuries CE). Conversely, unlike an older text such as one from the Dead Sea Scrolls, a text from this period is likely to not include the characters א, ו, and י for vocalisation purposes.

Within the context of NLP, Biblical Hebrew may be viewed as representative of a specific genre, register or domain. It is also important to note that, due to the Hebrew Bible's limited size, Biblical Hebrew contains solely about 9000 distinct words, 1500 of which are hapax legomena (Sáenz-Badillos, 1993).

2.2 LLMs/NLP and Hebrew

Several Large Language Models (LLMs) that focus on the Hebrew language have been proposed up to date. BERT's multilingual version, mBERT, features about 2000 Hebrew tokens (Devlin et al., 2019), and more recent Hebrew-specific models often use it as a baseline when evaluating their performance. In particular, there are several BERT-based Hebrew models, whose abilities in relation to the language's morphology have been specifically emphasised. HeBERT (Chriqui and Yahav, 2021) is trained on the Wikipedia and OSCAR datasets and released along with the sentiment analysis tool HebEMO. Its performance is noted to improve when sub-word rather than word-based tokenisation is performed. AlephBERTGimmel improves on an earlier model, AlephBERT (trained on Wikipedia, Twitter and OSCAR), in a variety of NLP tasks, including morphological segmentation and POS tagging, by simply increasing its vocabulary size from 50k to 128k tokens (Gueta et al., 2023).

The DictaBERT model (Shmidman et al., 2023) occupies the current state-of-the-art in a number

[1] *wayyiqtol* and *weqatal*

of tasks, including morphology-related ones and sentiment analysis. It is trained on 3B words, and its authors note that the masking of only whole words rather than word segments has improved its performance significantly. DictaBERT is released along with two sub-models, DictaBERT-morph and DictaBERT-seg, which specialise in the respective tasks of morphological annotation and the segmentation of particles such as prepositions and articles from words. For the purpose of this project, it is also worth mentioning BEREL, an additional model proposed by DictaBERT's research team, which is trained on Rabbinic rather than Modern Hebrew text (as found in the *Sefaria*[2] and *Dicta*[3]) online libraries. At the time of writing, the BEREL model is only available as a demo version.

Other notable Hebrew-related NLP tools include a challenge set, devised and tested by DictaBERT's authors, which includes 56k professionally annotated sentences composed around 12 pairs of homographs, a frequent phenomenon within the Hebrew language that interferes with the performance of automatic analysis (Shmidman et al., 2020).

3 Methods

3.1 Manual annotation

Within the framework of this study, manual annotation is applied to the extant manuscripts of the Book of Ben Sira, a poetically written text dating from the 2nd century BCE that features guidance concerning Jewish life and worship. The choice of text is based on several factors. First comes its presence among the Dead Sea Scrolls, which constitute the implied project's framework due to their large number and relatively recent discovery. To go further, the Book of Ben Sira has received attention not only in established but now partly outdated studies (Beentjes, 1997; Ben-Hayyim, 1973), but also in recent academic work that matches the standards of the modern digital era, notably Rey and Reymond (2024). It is also worth noting that the text has a high number of extant witnesses[4] and that its complex nature in terms of vocabulary, syntax and use of figurative language render its study generalisable to a large array of other Biblical Hebrew texts.

Annotation is performed by professionals in the field of Biblical and Jewish Studies. Word-level annotation is initially opted for and hypothesised to be of significant importance due to the Hebrew language's especially strong morphology. The utilised texts are manually collated into word-pair variants, and the variants are assigned a defined category. Two of the categories also contain subcategories, which are indicated if the word pair can be identified with them unambiguously. Currently, differentiation between the subcategories is not used in the automatic classification process. However, the subcategories' definitions and proportions are made use of in the derivation of synthetic data. Please see Table 1 for English examples of the least intuitive categories. Appendix A provides detailed information about the meaning of each category and subcategory.

Formatting conventions as outlined in École Biblique et Archéologique Française de Jérusalem (1955-1982), such as superscript dots over a letter or different types of brackets, are retained to denote degrees of uncertainty about a text's interpretation.

Table 1: English examples of the 'Morphological' and 'Lexical' variant categories.

Morphological		**Lexical**	
var1	var2	var1	var2
cat	the cat	cat	car
cat	and cat	cat	Kate
cat	my cat	cat	qat
cat	cats	cat	kitten
		cat	dog

Several subcategories may be indicated for a given word pair; for example, the variants הכל (*hakol*; 'the' + 'everything') and לכל (*lekol*; 'to' + 'everything') belong to the category 'morphological' and the subcategories 'determination' and 'preposition'. In contrast, as the same pair may not be indicated as belonging to more than one category in the process of automatic classification, the category deemed most representative is opted for.

Table 2 shows the distribution of annotated data per category and, where relevant, subcategory.

3.2 Synthetic data

Due to the described manually annotated data's limited size, data augmentation was undertaken in the face of generation of synthetic data. The issuing synthetic data is based on random words taken from the Dead Sea Scrolls and alternative witnesses of the present texts, as provided, annotated and aligned within the the Qumran-Digital library

[2] https://www.sefaria.org/

[3] https://library.dicta.org.il/

[4] Nine witnesses are used in the annotation: A, B, B margin, C, D, E, F (Geniza manuscripts), M (Masoretic manuscript) and 11Q5 (Dead Sea/Qumran manuscript).

Category	Count
Same	1735
Unclassifiable	659
Lexical	476
Synonym[a]	*104*
Metathesis	*16*
Phonetic affinity	*13*
Antonym	*9*
Letter interchange	*6*
Misspelling	*1*
Morphological	430
Orthographical	*145*
Grammatical	*116*
Coordination	*44*
Suffixed pronoun	*44*
Preposition	*36*
Singular/Plural	*14*
Determination	*11*
Masculine/Feminine	*3*
Plus/Minus	430
Inversion	28
Total	3758

[a] Note that not all entries within a category that contains subcategories are assigned a subcategory.

Table 2: Distribution of annotated data by category and subcategory (number of word pairs)

of the Göttingen Academy of Sciences and Humanities (Akademie der Wissenschaften zu Göttingen, 2021)[5]. All text was cleaned of reconstruction signs and tokenised into words, and all words were shuffled. The randomised sample consisted of just over 70k words. Words were deleted from the sample upon use.

Please refer to Appendix B for a description of the pipelines for data generation, which are elaborated based on each category and subcategory's definition as well as on observations derived from the annotated data, such as proportions of POS tags and average Levenshtein distances within a category. Imitation of more detailed characteristics, such as the distribution of Levenshtein distances, was opted against as robustness of the classifier models was sought. The majority of the data was derived through the application of hand-crafted rules on words from the described randomised dataset. Occasionally, external sources, such the Hebrew dictionary Milog[6] were also made use of. Finally, in the cases of the 'masculine/feminine' and 'suffixed pronoun' subcategories, a portion of the used word pairs were hard-coded.

Three synthetic datasets were composed, which differ by the number and proportion of entries by category and subcategory. 'Synthetic dataset 1' is of the same size as the annotated dataset and contains the same number of entries per category. Balance is sought for subcategories, even where the original data is highly unbalanced. 'Synthetic dataset 2' is such that when it is concatenated to the annotated dataset, 1000 entries per category are achieved. Once again, balance is sought for subcategories. The general logic of 'synthetic dataset 2' is followed for the composition of 'synthetic dataset 3', which, however, includes a significantly larger number of entries. When this dataset is concatenated to the annotated one, 10k entries per category are achieved. The original proportions of entries per subcategory are maintained. The smallest number of data points, associated with the 'misspelling' subcategory, comes at 64, whilst the annotated data features only a single entry of this subcategory.

3.3 Morphological labels

Two sets of morphological labels are associated to each word from the annotated pairs for use within classifier experiments: the professionally attributed labels present in the Qumran-Digital library and retrieved with the help of an API developed for the purpose by our team (henceforth, the 'gold standard') and labels assigned automatically with the DictaBERT-morph model (henceforth, the 'silver standard'). The gold standard labels are originally composed in German and feature the following information: 'lemma', 'word class', 'short definition', 'root designation', 'verb stem', 'verb tense', 'person', 'gender', 'number', 'state', 'augment', 'suffix person' and 'suffix number'.[7] The information present in the silver standard labels consists of each word's POS, gender, number, person, tense, prefixes and suffix.[8] The gold standard labels include a significantly higher number of categories, some

[5] Texts from caves 1, 2, 4 and 11 were used due to their large number, and the proportion of texts was doubled for cave 4 due to its significant size.

[6] https://milog.co.il/

[7] For detailed definitions of the gold standard categories, please refer to Appendix C

[8] For detailed definitions of the silver standard categories, please refer to Appendix D

of which are particularly conceived with Classical Hebrew in mind (e.g. 'state', 'augment') and are naturally absent from the DictaBERT model, which is based on Modern Hebrew. Similarly, some of the gold standard labels within comparable categories are perceptibly more domain-specific (e.g. word class 'name of a god', consecutive tenses). The sole occasion of higher specificity associated with the silver standard annotation is that coordinating and subordinating conjunctions form separate categories.

As gold standard labels are based on a limited number of professionally annotated texts, they are not derivable for a large portion of the synthetic data (and for potential future text that our variant classification may be applied to). Silver standard labels are therefore resorted to in relevant experimentation. In order to evaluate the latter's quality, we explored the derived labels[9] of readily mappable categories across the two standards, calculating the silver labels' accuracy with respect to the gold ones. The mapped categories were: the gold standard's 'word class' and the silver standard's 'POS'; and the two standards' 'person', 'gender' and 'number'. Please refer to Appendix E for the full mapping applied to POS tags. The 'dual' number, not present among the silver labels, was mapped to 'plural'. In turn, the '1,2,3' silver tag for 'person' was considered to always be correct. As the gold standard labels are based on a word's original context and can therefore have different values at different occurrences of the word, all possible values for a word were retrieved, and their frequencies were noted. The silver labels' accuracy was calculated in two discrete settings: a match against the most common gold label versus a match against any of the possible gold labels.

Please see Table 3 for the results of the performed evaluation. As expected due to the high number of categories, accuracy was by far lowest for POS tags. The most common mistakes consisted in auxiliaries or verbs being marked as nouns or proper nouns; and nouns being marked as proper nouns. Accuracy was very high (over 0.9) for 'number' and 'gender' in the second scenario. The most common mistakes for 'number' labels were 'plural' being marked as 'singular'; for 'gender' - 'masculine' being marked as 'feminine'; and for 'person' - '2nd' being marked as '3rd'.

[9]Not all words in the manuscripts were associated with professional labels, resulting in a 'null' value. In turn, the DictaBERT-morph model also accorded a 'null' value to a portion of the words. Out of all 4046 words in the annotated dataset, 3099 received gold standard labels and 3413 - silver standard ones.

	Ac (1)	Ac (2)	# g + s	# g	# s
POS	0.25	0.55	3079	3099	3413
Num	0.58	0.96	1826	2710	2803
Gen	0.85	0.92	1818	2706	2256
Per	0.75	0.84	396	857	557

Table 3: The accuracy of silver standard labels as compared to gold standard labels in two scenarios: a match with the first label in terms of frequency (1); and a match with any of the possible labels (2). The number of words annotated with labels is also included. Only the words with both types of labels were evaluated. Num: number; Gen: Gender; Per: Person.; g: gold; s: silver

3.4 Classifiers

A variety of multiclass classifier models are experimented with until maximal performance in terms of F1 value[10] is reached in the task of prediction of word-level variation in Biblical Hebrew text: these include Logistic Regression, Random Forests and Support Vector Machines (SVM)[11], as well as neural models based on DictaBERT as the current state-of-the-art in the Hebrew language. For the non-neural models, multiple parameters are explored in the context of a grid search, notably including different tokenisation methods.[12] The neural classifiers make use of the Python library *transformers*.[13] They are trained for 7 epochs[14], with 3 random seeds, and different train and evaluation batch sizes are tested. The experiments utilised nodes equipped with Intel Xeon Gold 5220 CPUs (2.20GHz, 18 cores) and 96 GiB of RAM, alongside two NVIDIA GeForce GTX 1080 Ti graphics cards.

The manually annotated dataset described in Section 3.1 is used in the training process, and synthetic datasets 1 to 3 (see Section 3.2) are added to it in both the neural and non-neural models, whilst evaluation is only performed on annotated data. In addition to the concatenated embeddings of the

[10]This balanced measure is opted for in an attempt to limit the extent of a model's individual shortcomings and thus render it robust.

[11]as available through the Python library *scikit-learn*: https://pypi.org/project/scikit-learn/

[12]CountVectorizer (unigrams), CountVectorizer (bigrams), TfidfVectorizer (unigrams), TfidfVectorizer (bigrams), TfidfVectorizer (2-5 grams)

[13]https://pypi.org/project/transformers/

[14]following observations about the point at which the results start to deteriorate

two words that comprise each pair, Levenshtein distance[15] and morphological characteristics (as per both the gold and silver standards elaborated in Section 3.3 for annotated data and the silver standard when synthetic data is included) are experimented with as features within the non-neural models. POS tags and other morphological information such as phi-features[16] have proven to be beneficial for morphologically rich languages in sentence-level tasks such as syntactic parsing (Marton et al., 2010; Collins et al., 1999; Tsarfaty and Sima'an, 2007), but to our knowledge no similar evaluation has been carried out at the word level. In contrast, research shows that neural models such as BERT do not benefit from explicit morphological labels, except in rare situations where the labels are of especially high quality[17] (Klemen et al., 2023). Finally, different hand-crafted rules per category are defined and set to increase the probability of the respective categories being predicted at inference.[18] These rules pertain to the categories' definitions and statistical observations based on the annotated data. A positive boolean value is attributed to the feature 'likely inversion' if 'word 1' and 'word 2' can be encountered reversed within four indices of the current index and the two words have Levenshtein distance of at least 2. 'Likely plus/minus' is marked positively when one of the words is empty, and 'likely unclassifiable' – when at least one of the words contains at least one square bracket. 'Likely morphological' is attributed when both words are present, do not contain square brackets and have a Levenshtein distance smaller than 2. A rule related to the lexical category is not developed due to the category's highest complexity as well as the fact that it is the sole remaining category.

Please refer to Figure 1 for an overview of the trained classifiers' input and output.

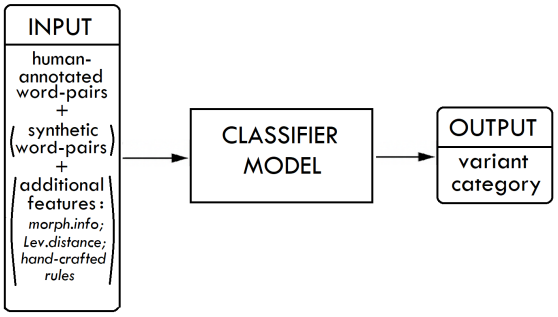

Figure 1: Classifier Input and Output

4 Results

Table 4 summarises the nature and performance of the key classifier models experimented with. For more detailed experimentation results, please refer to Appendix F. The strongest non-neural baseline models were Random Forests[19]. The optimal tokenisation technique was revealed to be TfidfVectorizer with character bigrams. Both gold and silver standard morphological labels were experimented with. In the case of the former, a setting where only the most frequent label per category was used proved to be the better choice. The highest performing non-neural model (which also performed best overall) had an F1 score of 0.80. It was trained on the annotated and 'synthetic 1' data and included silver standard POS tags and Levenshtein distance as features. The 'Unclassifiable' and 'Plus/Minus' categories presented a challenge, likely due to the fact that the former can be characterised with a missing variant and does not always include reconstruction signs at the word-level[20]. Very closely at second place came the model trained solely on annotated data and silver standard POS tags. The best non-neural models trained with the help of 'synthetic 2' and 'synthetic 3' data reached, respectively, F1 of 0.78 and 0.72. Curiously, silver standard morphological information led to consistently better performance than gold standard.

The DictaBERT-based neural classifiers achieved competitive but slightly lower results, with the exception of the classifier trained with the help of 'synthetic 3' data, whose performance was higher (0.74, against 0.72 for the non-neural models). The best F1 score was achieved by the model trained solely on annotated data (0.78), followed by the ones making use of 'synthetic 2' (0.77), 'synthetic 1' (0.76) and finally 'synthetic 3'. The neural models took a significant amount of time to train: between 6 min ('annotated') and 2 h 45 min ('annotated' plus

[15] A measure of cosine distance has also been experimented with but discarded as proven less effective.

[16] e.g. number, gender, person

[17] A setting that we cannot guarantee, especially in the presence of synthetic data.

[18] The impact that the rules have on a category's prediction in terms of an increase in percentage are determined based on a process of trial and error.

[19] The best parameters included 300 estimators for all data settings except annotated data plus 'synthetic dataset 2' (where the best value was 100).

[20] See Appendix G for a sample of predicted labels by this model versus their real counterparts.

Model	F1	Ac	Pr	Re
Base$_{An}$	0.67	0.67	0.67	0.67
Base$_{An+S1}$	0.68	0.68	0.69	0.68
Base$_{An+S2}$	0.67	0.65	0.70	0.75
Base$_{An+S3}$	0.66	0.66	0.74	0.61
Mod1$_{An}$	0.70	0.70	0.70	0.70
Mod2$_{An}$	0.74	0.75	0.75	0.74
Mod2$_{An+S1}$	0.72	0.72	0.73	0.72
Mod2$_{An+S2}$	0.68	0.67	0.70	0.67
Mod2$_{An+S3}$	0.69	0.69	0.80	0.69
Mod2+L$_{An}$	0.72	0.73	0.74	0.73
Mod2+L$_{An+S2}$	0.76	0.76	0.77	0.76
Mod2+L$_{An+S2}$	0.73	0.71	0.76	0.71
Mod2+L$_{An+S3}$	0.71	0.68	0.79	0.68
Mod2+R$_{An}$	**0.80**	**0.80**	**0.80**	**0.80**
Mod2+L+R$_{An+S1}$	**0.80**	**0.80**	**0.82**	**0.80**
Mod2+L+R$_{An+S2}$	**0.78**	**0.76**	**0.81**	**0.76**
Mod2+L+R$_{An+S3}$	**0.72**	**0.69**	**0.81**	**0.69**

Table 4: Non-neural classifiers.
An: annotated; S: synthetic
Base: Random Forests + TfidfVectorizer (char bigrams)
Mod1: Base + gold morphological labels 'word class'
Mod2: Base + silver morphological labels 'POS'
L: Levenshtein distance
R:'inversion', 'plus-minus', 'unclassifiable' and 'morphological' rules
Values are rounded to the second digit after the decimal point. The highest results per data setting are indicated in **bold**.

Models	F1	Ac	Pr	Re
NN$_{An.}$	**0.80**	**0.80**	**0.80**	**0.80**
NN$_{An.+S1}$	**0.80**	**0.80**	**0.82**	**0.80**
NN$_{An.+S2}$	**0.78**	**0.76**	**0.81**	**0.76**
NN$_{An.+S3}$	0.72	0.69	0.81	0.69
N$_{An.}$	0.78	0.78	0.79	0.78
N$_{An.+S1}$	0.76	0.76	0.77	0.76
N$_{An.+S2}$	0.77	0.77	0.78	0.77
N$_{An.+S3}$	**0.74**	**0.73**	**0.76**	**0.73**

Table 5: Best non-neural vs neural (DictaBERT-based) classifiers per data setting. NN: non-neural; N: neural

'synthetic 3'). The globally best train and evaluation batch sizes were 4 and 8, respectively. Table 5 summarises the best non-neural vs neural classifiers for each data setting.

5 Discussion

Whilst the highest performing model made use of an amount of synthetic data equal to the professionally annotated dataset, the applied data augmentation technique was not of significant benefit, in particular when neural models were concerned. Importantly, performance deteriorated perceptibly with the use of the largest synthetic dataset (10k data points per category), showing that further augmentation was unneeded. We conclude that the synthetic data failed to capture in sufficient detail the characteristics displayed by the annotated data. A hypothesis that remains to be tested is whether the models that include synthetic data have the quality of being more generalisable when different manuscripts (e.g. in terms of genre) are involved.

An analysis of the neural models' performance per category revealed that whilst some of the categories' results were improved by the use of synthetic data (e.g. 'Morpohological', 'Unclassifiable'), results for the 'Inversion' category were significantly weaker, reaching an F1 value of just around 0.33 (against 0.60 for models without synthetic data). As the order of word pairs is lost upon classifier training at the word level, we conclude that inverted word pairs within the utilised manuscripts exhibit characteristics that were not captured effectively by the synthetic data i.e. by a process of random inversion combined with an analysis of POS tags and Levenshtein distances.

Among the features used in non-neural classifiers, we note that silver standard POS tags performed significantly better than their gold standard counterparts, and despite a slight improvement, this remained the case even when the number of categories within the gold standard labels was reduced and they were made to match more closely the format of the silver ones. Possible explanations include a higher than expected quality of DictaBERT-based labels as well as their higher relevance to word-level analysis. The use of different combinations of morphological tags (e.g. number, gender, tense) in addition to POS tags led to varying performance that was always below that of POS tags when used in isolation. The Levenshtein distance between word pairs brought improvement of results

when synthetic data was involved, but only a minor one, possibly due to redundancy with the word representations themselves. It remains unclear why such improvement was not exhibited by models based only on annotated data. The utilised manual rules for separate categories had a significant positive impact. For instance, their use of indexing in determination of the 'Inversion' category helped overcome a serious limitation posed by word-level classification.

6 Conclusion and Future Work

The current project involves the derivation of a classifier model that predicts the category of word-pair variants as found in collated manuscript witnesses. The strongest model is a non-neural (Random Forest) one that makes use of professionally annotated data based on extant manuscripts of the Book of Ben Sira as well as automatically derived synthetic data. Additional features defined as useful at the classification process include hand-crafted rules per category, Levenshtein distance and POS tags. As professionally annotated morphological labels are only available for selected texts, this study used the opportunity to compare their performance to that of automatically derived labels by the state-of-the-art DictaBERT model. Curiously, the latter helped the classifier models achieve higher results.

Future plans pertaining to the authors' larger project include the use of the derived classifier to automatically annotate the word-level differences between multiple pairs of manuscripts, with a focus on the Dead Sea Scrolls. Consequently, the types and proportions of these differences are to be analysed statistically in view of their relevance in the determination of relationships between witnesses as present in established genealogical trees. Ultimately, automatisation of the process of tree generation will be sought.

The derived classification model and the pipelines for synthetic data generation are readily applicable to texts in Classical Hebrew, importantly including texts that have not received high engagement and benefited from professional morphological annotation as of now, such as translations into Hebrew of Deuterocanonical books. With some modifications, the developed tools (e.g. the pipeline for synthetic data generation) are applicable to additional languages and tasks within the general field of stemmatology and NLP-based work with manuscripts.

Limitations

The quality of morphological labels and, specifically, 'silver standard' ones, is not perfect, which can result in reduced performance of the trained classifiers. In turn, the derivation of synthetic data is also associated with limitations, such as the use of dictionaries that are markedly Modern-Hebrew. Also, the focus on word-level differences between textual witnesses as well as on word morphology, whilst hypothesised to serve as a reasonable proxy for the described texts' key characteristics, is not exhaustive. Alternative divisions of categories may also significantly alter the classification process; in particular, the latter would become a linear regression problem if variant differences are perceived as quantitative rather than qualitative, as they are in some stemmatological studies, e.g. Staalduine-Sulman (2005). Finally, the applied process of manual annotation based on a single text (the Book of Ben Sira) may also hold limited representativeness which, in turn, may be reflected in the developed classifier's applicability to other texts.

Ethics Statement

All utilised resources (such as gold standard morphological information) are publicly available. Manual annotation was performed in controlled conditions in the involved team's university headquarters. As an ancient language is concerned by the study, its importance as cultural heritage is acknowledged. The choices of language and texts are not based on religious or political motivations, and textual interpretation is not approached.

Acknowledgements

This work made major use of the manual collation and annotation of the examined witnesses of the Book of Ben Sira, performed by Davide D'Amico, postdoctoral reseracher in Research Centre Écritures.

References

Akademie der Wissenschaften zu Göttingen. 2021. Qumran-digital: Ein komplettes philologisches qumran-lexikon zum hebräischen und aramäischen. Accessed: 18.10.2022.

Pancratius C. Beentjes. 1997. *Book of Ben Sira in Hebrew*, volume 68 of *Vetus Testamentum, Supplements*. Brill.

Zeev Ben-Hayyim. 1973. *The Book of Ben Sira: Text, Concordance and an Analysis of the Vocabulary*. Academy of the Hebrew Language and the Shrine of the Book.

Bernard Cerquiglini. 1983. Eloge de la variante. *Langages*, (69):25–35.

Avihay Chriqui and Inbal Yahav. 2021. Hebert and hebemo: a hebrew bert model and a tool for polarity analysis and emotion recognition. *CoRR*, abs/2102.01909.

Michael Collins, Jan Hajic, Lance Ramshaw, and Christoph Tillmann. 1999. A statistical parser for czech. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *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. Association for Computational Linguistics.

Eylon Gueta, Avi Shmidman, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Joshua Guedalia, Moshe Koppel, Dan Bareket, Amit Seker, and Reut Tsarfaty. 2023. Large pre-trained models with extra-large vocabularies: A contrastive analysis of hebrew bert models and a new one to outperform them all. *Preprint*, arXiv:2211.15199.

Tuomas Heikkilä. 2023. *Computer-Assisted Stemmatology*. Routledge.

Katherine L.(Ed.) Jansen. 1990. The new philology. *Speculum*, 65(1).

Geoffrey Khan, Shmuel Bolozky, Steven E. Fassberg, Gary A. Rendsburg, Aaron D. Rubin, Ora R. Schwarzwald, and Tamar Zewi, editors. 2013. *Encyclopedia of Hebrew Language and Linguistics*, 1 edition, volume 1-4. Brill, Leiden.

Matej Klemen, Luka Krsnik, and Marko Robnik-Šikonja. 2023. Enhancing deep neural networks with morphological information. *Natural Language Engineering*, 29(2):360–385.

Yuval Marton, Nizar Habash, and Owen Rambow. 2010. Improving arabic dependency parsing with lexical and inflectional morphological features. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 13–21.

Ray Pritz. 2016. Biblical hebrew and modern hebrew: How much do they understand? *Journal of Biblical Text Research*, 38:203–219.

Miguel Pérez Fernández and John F. Elwolde. 1999. *An Introductory Grammar of Rabbinic Hebrew*. Brill.

Frédérique Michèle Rey and Eric Reymond. 2024. *A Critical Edition of the Hebrew Manuscripts of Ben Sira: With Translations and Philological Notes*. Brill, Leiden.

Philipp Roelli, editor. 2020. *Handbook of Stemmatology*. De Gruyter Reference. De Gruyter, Berlin/Boston.

William M. Schniedewind. 2013. *A Social History of Hebrew: Its Origins Through the Rabbinic Period*.

Avi Shmidman, Joshua Guedalia, Shaltiel Shmidman, Moshe Koppel, and Reut Tsarfaty. 2020. A novel challenge set for hebrew morphological disambiguation and diacritics restoration. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3316–3326, Online. Association for Computational Linguistics.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. Dictabert: A state-of-the-art bert suite for modern hebrew. *Preprint*, arXiv:2308.16687.

Eveline van Staalduine-Sulman. 2005. Vowels in the trees: The role of vocalisation in stemmatology. *Aramaic Studies*, 3:215–240.

Angel Sáenz-Badillos. 1993. *A History of the Hebrew Language*. Cambridge University Press, Cambridge.

Angela Taylor. 2019. A contrastive analysis between biblical and modern hebrew in the context of the book of ruth. *Journal of Biblical Studies*.

Reut Tsarfaty and Khalil Sima'an. 2007. Three-dimensional parametrization for parsing morphologically rich languages. In *Proceedings of the Tenth International Conference on Parsing Technologies*, pages 156–167.

École Biblique et Archéologique Française de Jérusalem, editor. 1955-1982. *Discoveries in the Judean Desert*. Clarendon Press, Oxford.

A Categories and Subcategories Used in Data Annotation

Same

This category, not represented in the classifier, is used to mark all word pairs in which the two items are identical or differ only in the presence or nature of diacritics or of symbols denoting a level of uncertainty about a letter's reading (e.g. רבים/רַבִּים; *ravim*; 'many' *masc. pl.*).

Plus/Minus

As for the purpose of this study, all texts are assumed to be of the same hierarchical level, the term 'plus/minus' is opted for as opposed to the commonly used in textual criticism 'addition' and 'omission'. The category is used for cases in which one of the two variants is missing.

Inversion

This category is used if 'word 1' in the given pair corresponds to 'word 2' in another pair found in close proximity in the manuscript and 'word 2' in the given pair corresponds to 'word 1' in the same closely situated pair. The corresponding words may be identical or feature minor differences, such as the addition of a coordinating conjunction or definite article.

Morphological

The difference implies the words' morphological features.

Determination

Only one of the variants features a definite article (e.g. יין/היין; *hayain/yain*; 'the wine'/'wine').

Orthographical

There is a spelling difference between the variants; in particular, the letters ה, ו, י and א may be added in one of the witnesses in order to aid vocalisation in a text that does not contain diacritical marks (e.g. פעל/פועל; *poal*; 'action').

Coordination

One of the variants includes the coordinating conjunction ו (e.g. לא/ולא; *velo/lo*; 'and no'/'no').

Preposition

One of the variants contains a prefixed preposition ל (*le*; 'to, towards'), ב (*be*; 'in') or כ (*ki*; 'as, like') or the two variants contain different ones (e.g. נגף/בנגף; *negaf/benegaf*; 'plague'/'in the plague').

Singular/Plural

There is a difference in number between the variants, which may be a textbook case of singular versus plural versions of a noun or adjective (e.g. מתן/מתנות; *mitnot/matan*; 'gifts'/'a gift') or involve higher formal complexity, such as in the case of suffixed possessive pronouns (e.g. דברך/דבריך; *dvareHa/dvarHa*; 'your words'/'your word').

Masculine/Feminine

There is a difference in gender between the variants (e.g. בנים/בנות; *banim/banot*; 'boys'/'girls'). Verb conjugations for different gender fall into the 'grammatical' rather than the 'masculine/feminine' subcategory.

Suffixed Pronoun

Only one of the words in the pair contains a suffixed possessive or direct object pronoun (e.g. רצון/רצונו; *retsono/ratson*; 'his will'/'will') or the two words contain different suffixed pronouns (e.g. לך/לה; *leHa/la*; 'to you'/'to her').

Grammatical

This is the broadest of the morphological subcategories and denotes different grammatical nature or function between the variants, such as different verb tense or form (e.g. יוסיף/מוסיף; *mosif/yosif*; 'to add' *participle* vs *imperfect*), different verb gender (e.g. היתה/היה); *haya/haita*; 'there was' *masc.* vs *fem.*), different part of speech (e.g. תקיף/תקף; *takaf/takif*; 'to attack'/'strong') or otherwise different words sharing the same root (e.g. שכיר/שוכר; *soHer/saHir*; 'tenant'/'hired worker'). A prefixed subordinating conjunction also implies this category (e.g. שיחפץ/יחפץ; *yaHpots/sheyaHpots*; 'he will desire' *masc. sing.* vs 'that he will desire'). Combinations of two or more grammatical differences may be involved (e.g. נואי/נאים; *naim/noei*; 'beautiful' *adj.* vs 'beauty, ornament' *noun in construct state*).

Lexical

The difference between the variants is at the lexical level.

Letter Interchange

There is a difference between the words in the pair pertaining to letters with high visual similarity (e.g. תדע/תרע; *tera/teda*; 'to harm'/'to know').

Phonetic Affinity

The two variants are pronounced in the same or similar way (e.g. מפליתו/מפלתו; *mafalto/maflito*; 'his defeat'/'his escape').

Metathesis

The difference involves replacement of letters which may have occurred as a result of language development, such as to facilitate pronunciation (e.g. חכמה/כחמה; *HoHema*; 'wisdom').

Misspelling

There is a mistake within the spelling of one of the words in a pair. This category is generally avoided, as the fact that a word is not readily recognisable does not automatically mean that it is a misspelling of another, more intuitive word.

Synonym

The two words in the pair are etymologically different but have the same or similar meaning, whether globally or in the given context (e.g. תמחה/תשכח; *tishkaH/timaHeh*; 'she forgot'/'she erased').

Antonym

The two words in a pair have opposite or contrasting meaning (e.g. וּלקח/וּתת; *vetet/velakaH*; 'he gave'/'he took'; חמס/חכם; *HaHam/Hamas*; 'wisdom'/'violence').

Unclassifiable

This category is used for instances where one or both of the variants are unidentifiable solely based on the given manuscript. Restored text with high uncertainty (i.e. marked with square brackets) is always attributed this category. Note that sometimes restored text encompasses multiple words, in which case square brackets are present only in the beginning and end of the group.

B Pipelines for Generation of Synthetic Data by Category and Subcategory

Plus/Minus

Either 'word 1' or 'word 2' (selected at random) is populated with a random word from the randomised Qumran sample until the desired number of entries is achieved.

Inversion

Random entries are generated where 'word 1' and 'word 2' with Levenshtein distance of at least 2 are taken from the randomised sample. Reversed versions of each pair are also composed. The matching entries are then organised so as to be either adjacent or nearly so following a distribution, close to the one in the manually annotated corpus.

Morphological

Determination

Words with POS 'verb', 'noun', 'adverb' of 'adjective' are taken from the randomised sample, ensuring that they do not already contain a prefix with the help of the DictaBERT-seg model. Word pairs are formed with the original words as 'word 1' and the identical words as preceded by the definite article (ה) as 'word 2'. The words in each pair are shuffled.

Orthographical

Words are taken from the randomised sample until the desired size is met. For half of them, random existing ו and י letters are removed. For the other half, ו, י and א (the last one in no more than 10% of cases) are added in random positions. It is ensured that initial ו used as a coordinating conjunction and י used as a possessive pronoun are not altered. The words in each pair are shuffled.

Coordination

Words are taken from the randomised sample, ensuring that no more than half of them are verbs and that they do not contain prefixed particles with the model DictaBERT-seg. Word pairs are formed from the original words and the same words preceded by the coordinating conjunction (ו). The words in each pair are shuffled.

Preposition

Words are taken from the randomised sample, ensuring that about half of them are nouns and 1/3 are pronouns and adpositions; and that they do not contain prefixed particles. A list of valid prepositions is defined: ב, ל, מ, כ. Word pairs are formed based on each random word in one of the following five scenarios: 'word 1' contains no prepositions and 'word 2' contains one preposition; 'word 1' contains no prepositions and 'word 2' contains two prepositions; 'word 1' and 'word 2' each contain a different preposition; 'word 1' contains a coordinating conjunction and 'word 2' contains a coordinating conjunction and a preposition; 'word 1' and 'word 2' each contain a coordinating conjunction and a different preposition. The words in each pair are shuffled.

Singular/Plural

Singular nouns are taken from the random sample. 'Word 1' is populated with the original words and 'word 2' with their pluralised versions (ים for masculine; ות or occasionally ת for feminine; any final הs are removed). The words in each pair are shuffled.

Masculine/Feminine

First, the base word pairs אח and אחות ('brother' and 'sister'), בן and בת ('son' and 'daughter') as well as several versions of them including pluralisation and randomly added prefixes and suffixes are taken. Then, several hard-coded common pairs

of masculine/feminine nouns are added[21]. Finally, adjectives without suffixes or final ה[22] are taken from the randomised sample and their gender is changed following hand-crafted rules: a final ה is added or removed to render a singular adjective respectively feminine or masculine; the endings ת (or תות) are replaced with ים and vice versa to render plurals feminine or masculine. All derived entries as well as the words in each pair are shuffled.

Suffixed Pronoun

As only transitive verbs can take suffixed pronouns, a number of verbs are taken from the sample and the transitive ones are filtered with the help of expert knowledge and ChatGPT[23]. Then, lists of all possible combinations of particles and pronouns suffixed to them are composed and pairs of them accounting for 10% of the desired subsample are added. For the rest of the entries, nouns are taken from the randomised sample and pairs are formed with the word as 'word 1' and the word plus a random pronoun[24] as 'word 2'; in 20% of cases, different random pronouns are added to both words. All derived entries as well as the words in each pair are shuffled.

Grammatical

Verbs are taken from the random sample to populate 'word 1'. For the population of 'word 2', the Modern Hebrew conjugation website Pealim[25], is web-crawled. In a first scenario, a different random conjugation of the same verb is taken. In a second scenario, a conjugation of an etymologically related word is used instead. In 20% of cases, 'word 1' is also changed with a related word, ensuring a variety in parts of speech. The words in each pair are shuffled.

Lexical

Letter Interchange

A dictionary of visually similar Hebrew letters is defined[26] and random words are sought that contain at least one of the implied letters. To form the word pairs, a random relevant letter is swapped from each word based on the dictionary. The words in each pair are shuffled.

Phonetic Affinity

A dictionary of phonetically similar Hebrew letters is defined[27] and random words are sought that contain at least one of the letters. To form the word pairs, a random relevant letter is swapped from each word based on the dictionary. The two words in each pair are shuffled.

Metathesis

Words are taken from the randomised sample, ensuring that at least 20% of them contain adjacent כ and ח letters. For the words containing said letters, their places are reversed. For the rest of the words, two random adjacent letters are swapped. The two words in each pair are shuffled.

Misspelling

Words are taken from the randomised sample. To form pairs, they are modified in one of the following scenarios: one of the word's letters is replaced by a random letter; two of the word's letters are replaced by random letters; one of the word's letters is deleted. In the case of words starting with a particle, the particle is not altered. The words in each pair are shuffled.

Synonyms and Antonyms

The Modern Hebrew online dictionary Milog[28] is crawled using the Python library *requests*[29]. Random one-word synonyms and antonyms of tokens from the randomised sample are taken until a defined number of entries are found for which at least either one synonym or one antonym exists. As diacritic signs are used in the dictionary but typically not in the Qumran texts, they are removed from a portion of the derived data. The words in each pair are shuffled. Antonyms were retrieved for about 1/4 of the sought random words, and synonyms - for a little over a half of them.

Unclassifiable

Words are taken from the randomised sample and square brackets are added at random positions within them in the following ways: either '[' or ']' is added to 'word 2' (2/3 of cases); both '[' and ']' (in this order) are added to 'word 1' (1/6 of cases);

[21] נערה/נער, מלכה/מלך, זקנה/זקן, חכמה/חכם, שופטת/שופט, סופרת/סופר, גברת/גיבור, רוֹעָה/רוֹעֶה, נביאה/נביא, כהנת/כהן

[22] as when non-vocalised, these adjectives look identical in the two genders

[23] as per GPT-4

[24] י, ד, ו, ה, נו, כם, כן, ם, ן

[25] https://www.pealim.com/

[26] כ and ב; נ and ג; ר and ד; ח and ה; ו and ן; ו and ז; ו and י; צ and ט; ך and ד; ך and ן; פ and ף; ס and ם; ח and ת

[27] א and ע; ש and ס; ט and ת; ב and ו; כ and ק

[28] https://milog.co.il/

[29] https://pypi.org/project/requests/

'[', ']' or both symbols are added to both words (1/6 of cases). The words in each pair are shuffled.

C Gold Standard Labels, Encountered in the Annotated Dataset

Lemma

The basic form or forms associated with a word. It may consist in the removal of signs denoting manuscript reconstruction and vocalisations; a division of the root and affixes; and the reduction to a default form in terms of number (singular), gender (masculine) or person (third).

Word class

The categories correspond roughly to conventional POS tags but involve higher specificity. For instance, pronouns and proper nouns are divided into subcategories (personal, question; name of a person, of a group of people, of a god, of a place). For relevant parts of speech, the different numbers, genders and persons form separate labels. 'Letter' is also a defined category. Some categories used in Universal Dependencies[30] (UD), such as 'punctuation', are not present.

Short definition

Translation or definition of the word in German.

Root designation

May take values of 'I', 'II', 'III', 'IV' or 'V'. It is related to the context-specific meaning of the root as indexed in the dictionary associated with the Qumran-Digital project[31].

Verb stem

The type or group of Hebrew verb implied (e.g. *hif'il*, *nif'al*, *pi'el*), which is often indicative of the verb's general meaning or aspect.

Verb tense

A verb's tense. May be 'imperfect', 'participle', 'perfect', 'imperative', 'construct infinitive', 'consecutive imperfect', 'consecutive perfect' or 'cohortative'.

Person

Used for applicable parts of speech. May be '1', '2' or '3'.

Gender

Used for applicable parts of speech. May be 'masculine', 'feminine' or 'common'.

Number

Used for applicable parts of speech. May be 'singular', 'plural' or 'dual'.

State

Used for applicable parts of speech. May be 'absolute', 'construct' (i.e. forming a genitive construction) or 'determination' (i.e. it includes a definite article or demonstrative pronoun).

Augment

Emphasises the subject's relationship to the action. The only detected value is 'energetic'

Suffix person

Designates the person implied by the suffix. May be 'singular' or 'plural'.

Suffix number

Designates the number implied by the suffix. May be '1', '2' or '3'.

D Silver Standard Labels, Encountered in the Annotated Dataset

POS

POS tags corresponding to UD conventions: ADP (adposition; preposition or postposition), ADV (adverb), AUX (auxiliary verb), CCONJ (coordinating conjunction), DET (determiner), INTJ (interjection), NOUN (common noun), NUM (numerical), PRON (pronoun), PROPN (proper noun), PUNCT (punctuation), SCONJ (subordinating conjunction), VERB (verb), X (not classified).

Gender

Used for applicable parts of speech. May be 'masculine', 'feminine' or 'masculine and feminine'.

Number

Used for applicable parts of speech. May be 'singular' or 'plural'.

Person

Used for applicable parts of speech. May be '1',

[30] https://universaldependencies.org/
[31] https://lexicon.qumran-digital.org/

‘2’, ‘3’ or ‘1, 2, 3’.

Tense

A verb’s tense. May be ‘future’, ‘past’, ‘present’ or ‘imperfect’.

Prefixes

Features a list value of the POS tags of any prefixes that the word contains.

Suffix

Features the POS tag of any suffixes that the word contains. Combinations of POS tags appear as a single predefined value (e.g. ADP_PRON).

E Gold vs Silver Standard POS Tags

Silver	**Gold**
VERB	verb
AUX	verb
NOUN	noun; noun masc; noun fem; common noun
ADP	preposition; object marker
CCONJ	conjunction
SCONJ	conjunction; relative particle
INTJ	negation; interjection
ADV	adverbial particle
PROPN	name of god; name of person; name of group; name of place; name of month; name of region
PRON	question pronoun, person; question pronoun, thing; demonstrative pronoun, masc sing; demonstrative pronoun, common plural; personal pronoun, 3 masc sing; personal pronoun, 2 masc sing; personal pronoun, 2 fem sing; personal pronoun, 3 fem sing; personal pronoun, 1 common sing; question pronoun, place
X	letter
ADJ	*None*
DET	*None*
NUM	*None*
PUNCT	*None*

Table 6: Mapping of silver to gold standard POS tags.

F Detailed Classifier Results

Model	**Data**	**F1**	**Ac**	**Pr**	**Re**
Base					
	An	0.67	0.67	0.67	0.67
	An + S1	0.68	0.68	0.69	0.68
	An + S2	0.67	0.65	0.70	0.75
	An + S3	0.66	0.66	0.74	0.61
Mod1 (1) (all)					
	An	0.66	0.66	0.67	0.66
Mod1 (2) (all)					
	An	0.66	0.67	0.67	0.66
Mod1 (2) (all but 'verb tense')					
	An	0.67	0.68	0.68	0.68
Mod1 (2) ('word class', 'number', 'verb stem', 'gender' and 'suffix-person')					
	An	0.67	0.67	0.67	0.67
Mod1 (2) ('word class', 'number')					
	An	0.66	0.67	0.67	0.67
Mod1 (2) ('word class')					
	An	0.69	0.70	0.70	0.70
Mod1 (2) 'word class, simplified')					
	An	0.70	0.70	0.70	0.70
Mod2 (all)					
	An	0.63	0.64	0.65	0.64
Mod2 (all but 'gender')					
	An	0.71	0.72	0.73	0.72
	An + S1	0.70	0.71	0.72	0.70
	An + S2	0.69	0.69	0.72	0.69
	An + S3	0.68	0.67	0.76	0.67
Mod2 ('POS')					
	An	0.74	0.75	0.75	0.74
	An + S1	0.72	0.72	0.73	0.72
	An + S2	0.68	0.67	0.70	0.67
	An + S3	0.69	0.69	0.80	0.69
Mod2 ('POS') + L					
	An	0.72	0.73	0.74	0.73
	An + S1	0.76	0.76	0.77	0.76
	An + S2	0.73	0.71	0.76	0.71
	An + S3	0.71	0.68	0.79	0.68

Mod2 ('POS') + R ('inversion')					
	An	0.75	0.75	0.76	0.75
Mod2 ('POS') + R ('plus-minus')					
	An	0.75	0.76	0.78	0.76
Mod2 ('POS') + R ('inversion', 'plus-minus')					
	An	0.75	0.76	0.78	0.76
Mod2 ('POS') + R ('inversion', 'plus-minus', 'unclassifiable')					
	An	0.79	0.79	0.80	0.80
Mod2 ('POS') + R ('inversion', 'plus-minus', 'unclassifiable', 'morphological')					
	An	0.80	0.80	0.80	0.80
Mod2 ('POS') + L + R ('inversion', 'plus-minus', 'unclassifiable', 'morphological')					
	An + S1	0.80	0.80	0.82	0.80
	An + S2	0.78	0.76	0.81	0.76
	An + S3	0.72	0.69	0.81	0.69
N					
	An	0.78	0.78	0.79	0.78
	An + S1	0.76	0.76	0.77	0.76
	An + S2	0.77	0.77	0.78	0.77
	An + S3	0.74	0.73	0.76	0.73

Table 7: Trained Classifiers
An: annotated; S: synthetic
Base: Random Forests + TfidfVectorizer (char bigrams)
Mod1: Base + gold morphological labels
Mod1 (1): all gold labels per feature considered
Mod1 (2): only the most frequent gold label per feature considered
Mod2: Base + silver morphological labels
L: Levenshtein distance
R: hand-crafted rules N: neural (DictaBERT-based) model
Values are rounded to the second digit after the decimal point.

G Sample Label Predictions

Word 1	Word 2	Real	Predicted
יי	יָוִי	Lexical	Lexical
	יחזיק	Plus_Minus	Unclassifiable
זהב	זהוב	Morphological	Morphological
	ערים	Plus_Minus	Plus_Minus
אם	ואם	Morphological	Morphological
זֹ֫ר	יד	Lexical	Lexical
נשים		Plus_Minus	Plus_Minus
כל	על	Lexical	Lexical
דִּ֫חד	תיחד	Lexical	Lexical
זה		Plus_Minus	Plus_Minus
מטיב	מטוב	Morphological	Morphological
בוראו	אל	Lexical	Lexical
ענה	(ושואַל)	Lexical	Unclassifiable
לא	ל[א	Unclassifiable	Unclassifiable
	כל	Plus_Minus	Plus_Minus
לִּ֫בב:	שאר	Lexical	Lexical
ואתפלל	הטיתי	Lexical	Lexical
ועל	על[ן	Unclassifiable	Unclassifiable
	ויט	Unclassifiable	Plus_Minus

Table 8: The real vs predicted labels of a sample of annotated word pairs as per the strongest achieved classifier model.