Identifying Translationese at the Word and Sub-word Level

Ehud Alexander Avner* Noam Ordan† Shuly Wintner‡

Abstract

We use text classification to distinguish automatically between original and translated texts in Hebrew, a morphologically complex language. To this end, we design several linguistically informed feature sets that capture word-level and sub-word-level (in particular, morphological) properties of Hebrew. Such features are abstract enough to allow for the development of accurate, robust classifiers, and they also lend themselves to linguistic interpretation. Careful evaluation shows that some of the classifiers we define are, indeed, highly accurate, and scale up nicely to domains that they were not trained on. In addition, analysis of the best features provides insight into the morphological properties of translated texts.

1 Introduction

Much research in Translation Studies suggests that the language of translated texts, often called *translationese*, exhibits different linguistic properties from the language of original, non-translated texts. The differences are so marked that automatic (machine learning based) classification techniques can distinguish between original and translated texts with high accuracy, and indeed, several translationese classifiers have been defined for a few European

*Department für Linguistik, Universität Potsdam, Germany
†Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen, Universität des Saarlandes, Germany
‡Department of Computer Science, University of Haifa, Israel

languages. In this work, we employ text classification for the investigation of translationese in a morphologically complex language, namely *Modern Hebrew*.

This work is, to the best of our knowledge, the first to address automatic identification of translationese in a *Semitic language*; we are also the first to train our classifiers on a corpus of twentieth-century literary texts. Another novelty of the present work is that we focus on morphological (and, more generally, sub-word) features. An advantage of morphological features is that they lend themselves to interpretation, i.e., to qualitative analysis, as they can potentially capture structural and stylistic differences between translated and original texts. Such differences are realized in more analytic languages (like English) on the token level.

We thus set out to design several feature sets that capture word-level and sub-word-level phenomena – specifically morphological properties – of Hebrew translationese, and thus focus on the linguistic information encoded in tokens and sub-tokens. As will be shown, using the output of a morphological analyzer does not suffice; more sophisticated feature engineering is called for. We present a novel approach to approximating Hebrew word structure by means of alphabet abstraction. This approach, when enhanced with morphosyntactic information (that is, part-of-speech tags), turns out to be one of the most accurate and scalable among the classifiers we define.

The main contribution of this work is the construction of accurate classifiers that identify Hebrew translationese and can scale up to domains they were not trained on. This is important not only theoretically; numerous studies have shown that Statistical Machine Translation (SMT) systems can benefit a great deal when knowledge of the direction of translation is incorporated into the language and translation models (Kurokawa, Goutte, & Isabelle, 2009; Lembersky, Ordan, & Wintner, 2011, 2012a, 2012b, 2013). Robust detection of translationese is thus highly relevant for SMT.

This is the first work to address the automatic identification of translationese in Hebrew (or any other Semitic language), and the first to focus on the morphological manifestation of translated texts' properties. We thus also contribute to better understanding of the transla-

tion product. In addition, we show that literary corpora are suitable for the development of scalable identification systems, and introduce a novel approach to approximating Hebrew word structure that might be applicable to other Semitic languages.

In the next section we survey existing work on the automatic identification of translationese. In Section 3 we introduce some relevant characteristics of Hebrew orthography and morphology. The experimental setup is described in detail in Section 4. The features we define, and the rationale for using them, are discussed in Section 5. We then list the results of several computational experiments in Section 6 and analyze them in Section 7. We conclude with directions for future research.

2 Related Work

The term *translationese* was coined by Gellerstam (1986), who compared texts originally written in Swedish with texts translated from English into Swedish, and concluded that the striking differences between them do not indicate poor translation but rather a *statistical phenomenon*, a systematic influence of the source language on the target language. More recent works have suggested that *all* translations, regardless of source and target language, share certain features characteristic to translated texts (Baker, 1993; Toury, 1995).

Baroni and Bernardini (2006) were the first to employ text classification to investigate and identify translationese. Their comparable corpus is a collection of articles from an Italian geopolitics journal; each article is treated as a data point, i.e., as a training instance. The source languages from which articles are translated are assumed to be mainly English, Arabic, French, Spanish, Russian, and other languages. Prior to the classification, the corpus is tagged and lemmatized, and proper names are replaced with a dynamic ID-marker.

The learning method they employ is Support Vector Machines (SVMs). They experiment with numerous feature sets: frequencies of unigrams, bigrams, and trigrams of words, lemmas, part-of-speech (POS) tags, and a mixed mode in which function words are left untouched in their surface form, while content words are substituted by their corresponding

POS tag. They also experiment with combinations of the single SVM classifiers trained on the aforementioned feature sets. Experiments are run using sixteen-fold cross-validation. Single-feature classifiers yield accuracy of at most 77.1% (word unigrams and mixed mode bigrams). Trigram models obtain 62.5%-71.5%. The worse classification model is POS unigrams with accuracy of 49.6%. The best classifier (86.7%) is a combination of five models: lemma unigrams and bigrams, unigrams and bigrams of the mixed representation, and POS trigrams. Good features include function words and morphosyntactic categories in general, and personal pronouns and adverbs in particular. The accuracy of some classifiers is said to outperform human judgment.

Kurokawa, Goutte, and Isabelle (2009) identify translationese in English and Canadian French, and show what impact their findings have on machine translation systems. The corpus they use is a large portion of the Canadian Hansard, transcripts of the Canadian parliament proceedings. Following Baroni and Bernardini (2006), they produce four different representations of the corpus: surface forms, lemmas, POS tags, and a mixed representation. They, too, train SVM classifiers using n-gram frequencies: 1–to–5-grams of POS tags and of the mixed representation, and 1–to–3-grams of surface forms and lemmas. Classification is performed on blocks of text of varying lengths and on sentences. Using ten-fold cross-validation, the best classification results are just below 90% accuracy (for blocks) and 77% (for sentences). Both these results are achieved by SVMs trained on word bigram frequencies. Classifiers focusing on linguistic patterns, i.e., POS tags and the mixed representation, yield around 85%. The authors find that "[g]lobally, the relationship between the feature representations is clear: word > lemma > mixed > POS," and that "there seems to be an optimal n-gram length: bigram[s] for words and lemmas, trigrams for POS and mixed" (p. 84). Finally, Kurokawa et al. show that the direction of translation has an impact on SMT: translation systems the direction of which is the same as the direction of the training data perform better than systems going in the opposite direction.

Ilisei, Inkpen, Pastor, and Mitkov (2010) train and test their system on a translated and non-translated Spanish technical and medical dataset. They set out to go beyond the practi-

cal purpose of developing a classifier for translationese and "explore the characteristic [universal] features which most influence the translated language" (p. 504). Specifically, they are interested in the contribution of features designed to capture the *simplification hypothesis* (Blum-Kulka & Levenston, 1983; cf. also Baker, 1993) to the identification of translationese. According to this hypothesis, outputs of translators are less complex in terms of grammar, vocabulary, etc., than the source texts they render. Ilisei et al. propose various such 'simplification features': average sentence length, sentence depth (i.e., parse-tree depth), and lexical richness (type-token ratio), among other features. They compare several classification algorithms; the classifiers are trained on POS frequencies, including and excluding the simplification features. They find that removing the simplification features leads to decreased accuracy, and lexical richness is found to be the most informative feature. The best accuracy, 97.62%, is obtained by an SVM classifier. Ilisei and Inkpen (2011) apply similar methods to Romanian newspaper articles and obtain similar results.

Popescu (2011) studies English translationese at the character level. His corpus consists of 214 book-length literary works, most of them from the nineteenth century. The subcorpus of original English contains 108 works written by British and American authors; the translation subcorpus contains 76 works translated from French and 30 works translated from German. In the present work, we, too, train our classifiers on a literary corpus. Unlike Popescu, we strictly use *twentieth-century* literature.

The features Popescu extracts are simply character 5-grams, irrespective of word and sentence boundaries. Classification is performed on the book level (i.e., the training and testing instances are *complete* books), using SVMs and ten-fold cross-validation, and achieves virtually 100% accuracy. However, when the SVM is trained on British English original texts and on translations from French, but tested on American English and on translations from German, the accuracy drops to 45.83%, implying that the classifiers are overfitting. Popescu repeats the previous experiment, this time eliminating from the feature space all 5-grams that the French *original* texts and their translated counterparts share. The accuracy obtained this time is 77.08%. The advantages of the character n-gram approach are obvious: it

is language-independent, does not presuppose any language processing tools, and seems to promise relatively high classification accuracy. We hypothesize that character n-grams capture morphological features, and that such features could also be captured with n-grams shorter than 5.

Koppel and Ordan (2011) identify translationese, but also detect the source language of translated texts. They work on English translated from several languages (Finnish, French, German, Italian, and Spanish), using the EUROPARL corpus, which records the proceedings of the European Parliament (Koehn, 2005). The feature set in all their experiments is a list of 300 function words, and the learning method is Bayesian logistic regression. Training is done on 2,000 chunks of 2,000 words each, half of which are original English, the other half translated English (where each source language constitutes exactly one fifth of the translated data). Using ten-fold cross-validation, they identify translationese with 96.7% accuracy. The source language is correctly classified in 92.7% of the chunks. In addition, they train a classifier on EUROPARL and test it on a different corpus containing newspaper articles in original English and in English translated from Greek, Hebrew, and Korean. This classifier obtains 64.8% accuracy. In the opposite setting, i.e., when training on newspaper articles and testing on EUROPARL, the result is worse – namely 58.8%. We adopt their setup, working with 2000-word chunks. We, too, test our classifiers on datasets very different from the ones they are trained on. Unlike Koppel and Ordan, one of our goals is to find meaningful feature sets that are able to scale up to out-of-domain corpora.

Most recently, Volansky, Ordan, and Wintner (forthcoming) distinguish between original English and English translated from ten source languages (Danish, Dutch, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish). They, too, use parts of EUROPARL as their corpus: 4 million tokens of original English and 400,000 tokens from each of the source languages are partitioned into chunks of 2,000 tokens which are then used as training instances. The classification algorithm is SVM; testing is done using ten-fold cross-validation. Similar to Ilisei et al. (2010), Volansky et al.'s objective goes beyond the development of a working identification model. They set out to test several hypotheses – e.g., the

simplification hypothesis mentioned above – that have been purposed by translation scholars as translation universals. To this end, they define several feature sets that reflect these universals. Following Popescu (2011), they train classifiers using frequencies of character *n*-grams. In addition, they employ a precompiled list of prefixes and suffixes as a feature set approximating English morphological structure. The former classifier achieves 100% accuracy, the latter 80%.

Like Volansky et al. and Ilisei et al. (2010), our objective is not only to design accurate identification models, but also to explore computationally the properties of translated texts that distinguish them from original ones. However, in contrast to Volansky et al. and Ilisei et al., we work on a morphologically rich language with nonconcatenative morphology, and focus exactly on the word-level and sub-word-level features that cannot be investigated in more analytic languages such as English. Like Koppel and Ordan (2011), and unlike all other works mentioned above, we test our classifiers on datasets from domains different from the one they are trained on.

3 Hebrew Orthography and Morphology

The Hebrew alphabet is a 22-letter *abjad* (Daniels, 1997) for which two main standards exist: the *full script*, in which vocalization diacritics decorate words, thereby explicating all vowels, and the *lacking script*, in which these diacritics are missing, and the two letters *w* and *i* are occasionally added to represent some, but not all, of the vowels which would otherwise be represented by diacritics.[1] The overwhelming majority of Modern Hebrew texts – and all the texts our classifiers are trained and tested on – are written in the *lacking* variant. In this script, most of the five Hebrew vowels are left underspecified: /*e*/ and /*a*/ are usually not explicated (when they are, they are typically realized by the characters *a* and *h*); /*o*/ and /*u*/, when specified, are realized by the same character, *w*, which is also used for the consonant /*v*/. Similarly, the single character *i* is used both for the vowel /*i*/ (when it is specified) and

[1] For the sake of readability, a straightforward ASCII transliteration of Hebrew is used in this study. The characters, in Hebrew alphabetical order, are *abgdhwzxTiklmnsypcqr$t*.

for the consonant */y/*. The four characters which represent (some of) the vowels – *a, h, w, y* – are traditionally known as *matres lectionis*; they also play a significant role in Hebrew derivational morphology.

Many particles are realized as prefixes attached to the words immediately following them. These include the definite article *h*, the coordinating conjunction *w* ("and"), four of the most frequent prepositions – *b* ("in"), *k* ("as"), *l* ("to"), and *m* ("from") – and subordinating conjunctions, such as *k$* ("when") and *$* ("that/which"). When one of the prepositions *b*, *k*, or *l* precedes the definite article *h*, the latter is assimilated with the prefixing preposition and the resulting surface form becomes ambiguous with respect to definiteness.

Hebrew has a rich, partly nonconcatenative morphology. Derivational processes are based on a *root-and-pattern* system; inflectional processes are mainly carried out by suffixation, but also involve prefixes, circumfixes, and pattern shifts.

An example of the root-and-pattern mechanism are the seven Hebrew *binyanim*, i.e., the verbal patterns. Each pattern (*binyan*) is traditionally associated with a certain (vague, and thus not always predictable) meaning. For example, the *Hif'il* pattern productively generates causative variants of verbs; similarly (and, to a lesser extent), *Hitpa'el* is used for reflexives; three patterns systematically express the passive voice of three counterpart patterns, namely, *Nif'al* (vs. *Pa'al*), *Pu'al* (vs. *Pi'el*), and *Huf'al* (vs. *Hif'il*). Consider the three-letter root *k.t.b*, broadly denoting the notion of writing. When combined with the *Pa'al* pattern CCC,[2] this root yields the form *ktb* ("write"); when combined with the *Nif'al* pattern, nCCC, traditionally the passive counterpart of *Pa'al*, it yields *nktb* ("being written"); when combined with the *Hif'il* pattern, hCCiC, traditionally denoting causativization, it yields *hktib* ("dictate"). These morphological patterns are mechanisms for expressing constructions that require syntactic or lexical solutions in other languages. Similarly, a root can be combined with nominal patterns, for instance, *hktbh* ("dictation") is the result of combining the *k.t.b* root with the nominal pattern hCCCh which typically produces nominalized forms of the verbal pattern *Hif'il*.

[2]The C's in the pattern represent the slots for the three consonants of the root.

Verbs inflect for number, gender (masculine and feminine), person, and tense. *kwtbt*, for instance, is the present tense feminine singular realization (underspecified for person) of the *Pa'al* verbal pattern. The Hebrew tense system is relatively simple, with three tenses and no aspectual distinctions. Note that the present tense is actually a *participle* form that can also be used as an adjective or a noun (akin to *-ing* forms in English). Nouns inflect for number, adjectives for number and gender, numerals for gender.

Nouns, adjectives, participles, numerals, and quantifiers have two morphologically (and often phonologically) distinct forms: the unmarked *absolute* state and the *construct* state. The latter is used, in the case of nouns, adjectives, and participles, for the construction of compounds. It is also involved, in the case of nouns, in possessor-possessed constructions (i.e., noun compounds are, in point of fact, lexicalized possessive constructions). For example, *$mlh* ("dress," absolute state) vs. *$mlt* ("dress," construct state), as in *$mlt klh* ("dress," construct + "bride," absolute → "wedding dress," but also "a dress of a bride"). The fact that, in the *lacking script*, approximately half of the construct forms are orthographically identical to the absolute forms adds substantially to the ambiguity of Hebrew word forms.

There are several ways to express possessiveness in Hebrew, one of which is by attaching pronominal suffixes that inflect for number, gender, and person. The base form for these constructions is the construct state. For example, the first person singular possessive suffix *i* can be attached to *$mlt* ("dress", construct) to yield *$mlti* ("my dress"). The other possessive constructions go beyond the word level and involve the preposition *$l* ("of").

The morphological complexity, the deficient orthography, and the affixation of frequent particles bring about a system that produces highly ambiguous texts: "First, the first and last few characters of each token may be either part of the stem or bound morphemes (prefixes or suffixes). Second, the lack of explicitly marked vowels yields many homographs" (Fabri, Gasser, Habash, Kiraz, & Wintner, 2014). Hence, word segmentation is not straightforward, POS tagging is "a much messier task [...] than in other languages, such as English" (Koppel, Mughaz, & Akiva, 2006), and automatic morphological analysis is an immensely complex enterprise.

4 Experimental Setup

This is a corpus-based study; we use several corpora, which we automatically pre-process, to train and test machine learning based classifiers. We now explain the experimental setup and our methodology in more detail.

4.1 Corpus Design

The main corpus used in this study is a subset of a corpus compiled by Jason Perry[3] with the purpose of comparing translated and non-translated Hebrew texts. It is a monolingual comparable corpus that consists of first chapters of books published in Hebrew in the last decade (usually only the first chapter from each book). Perry downloaded the texts from a public Internet site aimed at exposing readers to newly published books.[4]

The corpus is annotated with the following metadata information: author name, book title, and source language. We add the following fields: translator name, genre (prose, play, verse, children's literature, etc.), and the author's year of birth. To allow for a better comparability, we restrict our training data to texts written by authors born after 1900, to English as the source language (of the translated texts), and to prose as the genre.

The subcorpus of original Hebrew literature (henceforth O_{heb}) contains 176 book chapters by 156 authors. The translation subcorpus (T_{en}) contains 128 book chapters by 123 authors translated from English by 81 translators. Each subcorpus contains exactly 600,000 tokens.[5] O_{heb} and T_{en} are the corpora we train our classifiers with; henceforth, we occasionally refer to them as the *training data*; together, they constitute the *InC* [in-corpus] experimental scenario introduced in Section 6.

Recall that most work on the identification of translationese has been carried out on corpora containing data from restricted domains, e.g., geopolitics (Baroni & Bernardini, 2006), parliament proceedings (Kurokawa et al., 2009; Koppel & Ordan, 2011; Volansky et al., forth-

[3] `http://hebrewcorpus.nmelrc.org/`, accessed 25 June 2013.
[4] `http://text.org.il/`.
[5] In this study, punctuation marks are counted as tokens.

coming), or technical and medical data (Ilisei et al., 2010). We use a corpus containing twentieth-century literary texts in this study, first and foremost, because we are not aware of any other large-scale comparable Hebrew corpus containing texts from domains such as the above. In fact, we are not aware of any other large-scale comparable Hebrew corpus of any domain.

There are, moreover, several benefits and interesting aspects to using a literary corpus: 1. The quality of translation is arguably very high. Not only can literary translators be assumed to be very competent translators, the common practice is that literary translations pass through an editorial cycle (copyediting, proofreading) before actually being published. 2. The multitude of authors and translators in our training data ensures that the classifiers do not learn to identify a *specific* author or translator but rather the phenomenon of Hebrew translationese. 3. A corpus of contemporary literature could be easily expanded for future research: in the age of the Internet the majority of publishers make excerpts of newly published books available online. 4. Metadata, such as the source language of the text, the birth date of the author, or the name of the translator, can normally be extracted with relative ease. 5. Identifying translationese by training on a corpus containing twentieth-century literature affords us an opportunity to explore a domain which very little work has been done on (one exception is Popescu (2011), whose corpus, however, consists of *nineteenth-century* literature). In fact, classifying literary translations is probably a harder task than classifying other genres, both because of the diversity of the texts and because much effort is invested in the translation of literary works, and more freedom is given to the translator to render the text as similar as possible to original writing. This is in contrast to more "technical" translations, which are often done under strict deadlines, resulting in more source-influenced, less fluent translations. Indeed, addressing a different but related task, namely, source language detection, Lynch and Vogel (2012), who train and test their models on nineteenth-century literary texts, state that they believe that literary translations "will pose a greater challenge [...] than the EUROPARL corpus, which is more homogeneous in style" (p. 778). 6. Finally, classifiers trained on a literary corpus might be able to scale up to scenarios which they are not trained

on. Baroni and Bernardini (2006), referring to their corpus, state that "[a] drawback of having a very uniform, very comparable corpus is that the results of our experiment may be true only for the specific genre and domain under analysis" (p. 264). We conjecture that a corpus of contemporary literature, being more heterogeneous than other closed-domain corpora, is suitable for the development of robust identification models.

We use additional datasets in order to check to what extent our classifiers scale up to scenarios they are not trained on. First, we test whether our models can predict translationese within the same domain (literature), but on texts translated from a different source language (French). Secondly, we test how well the classifiers predict translationese in a different domain, but on texts translated from the same source language as the training data (i.e., English). This last task is notoriously difficult (Argamon, 2011). None of the works discussed in Section 2, with the exception of Koppel and Ordan's (2011), test their systems on a domain different from the one their systems are trained on.[6]

In-Domain Corpus We construct a small corpus containing book chapters translated from French, rather than from English, extracted from the preliminary full corpus compiled by Perry. We refer to this corpus as the *in-domain* dataset (InD_{fr}). It includes 18 book chapters by 17 authors translated by 13 translators, totaling 60,000 tokens.

Out-of-Domain Corpora We use two additional small datasets: one containing journal and newspaper articles dealing with social science topics, often in a popular science style, the other containing journal and newspaper articles from the economics domain. We refer to these corpora as the *out-of-domain* datasets (OoD-soc[ial], OoD-eco[nomics]). OoD-soc consists of 33 translated articles and 26 original articles, and OoD-eco of 32 translated and 39 original. The number of authors and translators in these datasets is unknown; however, since the texts come from several different newspapers and journals, it is safe to assume that no one author (or translator) is overrepresented. Each of the OoD datasets contains 26,000 tokens of texts translated from English and 26,000 tokens of texts originally written in Hebrew.

[6] Popescu (2011) tests his model on texts translated from a different source language but in the *same* domain.

4.2 Morphological Analysis and Chunking

After applying a minimal cleaning script to the data, the corpora are first tokenized and then morphologically analyzed using the MILA tools (Yona & Wintner, 2008; Itai & Wintner, 2008). The *morphological analyzer* is a rule-based computational implementation of the inflectional morphology of Modern Hebrew, based on a lexicon of almost 30,000 lemmas. The morphological processor produces, for each token, its POS category.[7] Then, according to the POS, several other properties are specified. For verbs, e.g., these properties include *binyan* (verbal pattern, cf. Section 3), *gender*, *number*, *person*, and *tense*. In addition, the morphological analyzer segments tokens by specifying the sequence of affix particles attached to them, as well as the form and function of these affixes. As an example, Figure 1 depicts the output of the morphological processor on the word forms *wk$htxlti* וכשהתחלתי ("and when I began") and *sprihm* ספריהם ("their books"). Observe that in the first example, two prefixes are identified (ו, *w* "and", followed by כש, *k$* "when"), followed by the lemma *htxil* "begin". Then, the main POS is listed (verb), followed by a sequence of morphological features. The second example shows also a suffix, יהם (*ihm*), denoting a possessive pronoun in third person, masculine, plural ("their"). We come back to these examples in Section 5.2 below.

The output of the analyzer is disambiguated using the tagger of Bar-Haim, Sima'an, and Winter (2008): This is a stochastic tagger, trained on newspaper articles, and it ranks the analyses produced by the analyzer by assigning a score to each analysis (typically, '1.0' for the correct analysis, '0.0' for the incorrect ones). Unfortunately, the tagger is unable to always produce a unique top-ranked candidate; in cases where the tagger returns more than one optimal candidate, we simply pick the optimal result appearing first in the output. The reported accuracy of the POS tagger is 88.5%, but this evaluation is based on cross-validation experiments. As is well-known, out-of-domain evaluation of similar tasks usually reveals poorer performance. This is indeed our observation: on our corpus, the accuracy of tagging

[7] The tagset includes 25 tags: adjective, adverb, conjunction, copula, existential, foreign, interjection, interrogative, modal, MWE (multi word expression), negation, noun, numberExpression, numeral, participle, preposition, pronoun, properName, punctuation, quantifier, title, unknown, url, verb, and wordPrefix.

```
<token id="1" surface="וכשהתחלתי">
    <analysis id="1" score="1.0">
        <prefix function="conjunction" id="1" surface="ו" />
        <prefix function="temporalSubConj" id="2" surface="כש" />
        <base dottedLexiconItem="הִתְחִיל" lexiconItem="התחיל"
            lexiconPointer="17108" transliteratedLexiconItem="htxil">
            <verb binyan="Hif'il" gender="masculine and feminine" number="singular"
                person="1" register="formal" root="תחל" spelling="standard" tense="past" />
        </base>
    </analysis>
</token>

<token id="2" surface="ספריהם">
    <analysis id="2" score="0.5">
        <base dottedLexiconItem="סֵפֶר" lexiconItem="ספר" lexiconPointer="6651"
            transliteratedLexiconItem="spr">
            <noun gender="masculine" number="plural" register="formal"
                spelling="standard" />
        </base>
        <suffix function="possessive" gender="masculine" number="plural" person="3" />
    </analysis>
</token>
```

Figure 1: Output of the morphological analyzer on the tokens *wk$htxlti* וכשהתחלתי ("and when I began") and *sprihm* ספריהם ("their books").

seems to be lower, although we do not have precise data. In particular, the tagger often fails to distinguish between verbal analyses that differ in the *binyan* only. We also do not have data on the accuracy of the tagger on identifying any specific feature, but a different Hebrew tagger (Lembersky, Shacham, & Wintner, 2014), reporting similar overall accuracy on the same test set, reports over 92% accuracy on main POS, around 95% for number, gender, and person, over 98% on tense, etc.

Once a corpus is tokenized, analyzed, and tagged, it is partitioned into chunks, each containing 2,000 tokens;[8] there is no correlation between the number of chunks we extract from a corpus and the number of texts (i.e., chapters or articles) this corpus contains. That is to say, each chunk contains exactly 2,000 tokens, regardless of chapter/article and sentence boundaries. Since the main objective of this work is to observe word-level and sub-word-level phenomena in general, and to learn from morphological features packaged in single words in particular, we do not alter the size of instances; we treat each corpus (translated and non-translated) as a single continuous stream of data. We believe that 2,000-token chunks

[8] Punctuation marks count as tokens.

strike a balance between having enough chunks per corpus, on the one hand, and having big enough chunks to avoid problems of sparsity for certain rare word-level and sub-word-level features, on the other hand. Since none of our classifiers goes beyond the word level, sentence boundaries are irrelevant.

Table 1 summarizes the properties of the corpora used in the study.

	Chunks	Tokens	Texts	Authors	Translators	Split
O_{heb}	300	600,000	176	156	_____	All original
T_{en}	300	600,000	128	123	81	All translated from English
InD_{fr}	30	60,000	18	17	13	All translated from French
OoD-soc	26	52,000	59	_____	_____	50% orig., 50% trans. from Eng.
OoD-eco	26	52,000	71	_____	_____	50% orig., 50% trans. from Eng.

Table 1: The literary corpora: training (O_{heb} and T_{en}) and test (InD_{fr}, translated from French); and the *out-of-domain* test corpora.

4.3 Methodology

The core of our experimental methodology is the development of *classifiers* that can automatically distinguish between instances belonging to different classes (in our case, there are only two classes: translated and non-translated texts). The classifiers are trained on a corpus containing *training data*, that is, instances of the classes to be distinguished, each labeled a priori as belonging to one of the classes. Each of these instances is represented as a *feature vector*, a set of numeric attributes designed by the developers of the classifier to capture certain characteristics of the classes. The values of these features are extracted from the training instances (e.g., frequencies of certain words in an instance; see next section). During training, the classifiers learn to distinguish between the labeled instances, thereby assigning different weights to the features. A trained classifier can then be applied to unseen test instances and determine their class. If the features selected to represent the instances are meaningful, the classifier should be accurate when applied to test data.

Such methodologies have been extensively and successfully used for the automatic classification of texts according to, e.g., topic or genre (Sebastiani, 2002). They have been simi-

larly used for automatic *author attribution*, i.e., "inferring characteristics of the author from the characteristics of documents written by that author" (Juola, 2006, p. 233), for example, for identifying authors of newspaper articles (Diederich, Kindermann, Leopold, & Paass, 2003), or for determining the gender of a document's author (Koppel, Argamon, & Shimoni, 2002).

Support Vector Machine (SVM) is the classification algorithm employed in all our experiments. SVMs "probably represent the most successful technology for text categorization today" (Witten & Frank, 2005, p. 340), and indeed, SVMs have been widely and successfully used for identifying translationese (e.g., Baroni & Bernardini, 2006; Kurokawa et al., 2009; Popescu, 2011; Volansky et al., forthcoming). Specifically, we apply the Sequential Minimal Optimization algorithm (SMO) for training SVMs (Platt, 1999), using the default linear kernel, as implemented in the Weka machine learning toolkit (Hall et al., 2009).

All the identification models are trained and tested on the corpus containing O_{heb} and T_{en} in a ten-fold cross-validation procedure (we later refer to this experimental scenario as the *InC* [in-corpus] scenario). The obtained SVM classifiers are then also tested on the three additional datasets discussed above (InD_{fr}, OoD-soc, and OoD-eco). For all the experiments we report *accuracy*, namely the percentage of text chunks the classifier correctly classifies. In Section 7 we analyze the resulting classifiers, exploiting the values of the *weights* assigned by SVMs to the features used for classification.

5 Feature Design

The essence of text classification is the design of the feature vectors by which the text data are represented. As we do not go beyond the word level in this study, we design several feature sets aimed at capturing linguistic – specifically morphological – characteristics of surface tokens and sub-tokens. In this section we describe and motivate these feature sets.

5.1 Token-based Features

We use two different kinds of token-based features: *word unigrams* extracted from the training data, and a precompiled list of *function words* extracted from external corpora. In both settings, a list of tokens constitutes the feature vector representing a chunk, and feature values are the frequencies of these tokens in the chunk.

Word unigrams We compile a list of all the words in the training data, i.e., in the union of T_{en} and O_{heb} (excluding punctuation), and use each word as a feature. Like Volansky et al. (forthcoming), we treat this experiment as a sanity check, since, being highly content-dependent, this feature set is expected to yield very good classification results when tested on the training corpus in a ten-fold cross-validation scenario, but not to scale up to external domains.

Function Words Since Mosteller and Wallace's seminal work on the *Federalist Papers* (1964), function words have been extensively and successfully used in text classification. This approach for feature design has also been proven to be instrumental in identifying translationese, albeit not very scalable (Koppel & Ordan, 2011). These words "carry little meaning by themselves [...] but [...] define relationships of syntactic or semantic functions between other ('content') words in the sentence [... they] are therefore largely topic-independent and may serve as useful indicators of an author's preferred way to express broad concepts" (Juola, 2006, p. 242). Being highly frequent, these words exist in every chunk of text, regardless of its size, and since they are so frequent, it is safe to assume that more often than not text producers do not control the use of these words, i.e., do not select them consciously.

Unlike English, however, Hebrew text tokens often contain more than one lexical item, and many typical function words, such as prepositions, are concatenated to other words belonging to other parts of speech (cf. Section 3 above). Hence, closed sets containing several hundred function words, like the ones used for English, cannot be compiled for Hebrew. The list we use in this study contains Hebrew words belong-

ing to the following categories: quantifiers, pronouns, prepositions, negation words, interrogative markers, existentials, copulas, and conjunctions. It contains all possible inflections for each word – and only those surface forms that appear at least once in a collection of six large external Hebrew corpora. The list (which is available from MILA)[9] contains 7,450 items. Due to the morphological and orthographic challenges Hebrew poses (e.g., the fact that many function words are realized as affixes), classifiers based on function words are not expected to perform on our data as well as they do on English texts.

5.2 Features that Reflect Morphological Aspects

Since we are interested in investigating the morphological aspects of translationese, we define a set of features that reflect such information. To this end, we use the output of the morphological processor mentioned above (Section 4.2). Based on processor's output (cf. Figure 1), we define the following feature sets:

POS While POS tags may be considered syntactic rather than morphological features, we mainly employ them, as will be described below (Section 5.5), in order to enhance the performance and sophistication of other feature sets. We also use them, like Ilisei et al. (2010) do, as a baseline for testing the contribution of other features, in our case the 'pure' morphological features; i.e., we first train a classifier based solely on the 25 POS tags in the tagset, and then test this classifier with each of the morphological features added to it, and also with combinations thereof. This should give us a good indication of the contribution made by each morphological feature. For example, in Figure 1, the value of POS is *verb* in the first example, and *noun* in the second.

BINYAN The features in this category are the seven Hebrew verbal patterns, the *binyanim* (cf. Section 3 above). Since the verbal patterns have no counterpart in English, the source language of our study, we expect the frequencies of at least some of them to

[9] `http://mila.cs.technion.ac.il`

differ between original and translated texts. In Figure 1, the value of BINYAN in the first example is *Hif'il*.

STATUS The two features in this category are applicable to nouns, adjectives, participles, numerals, and quantifiers: the features *construct* and *absolute* reflect the construct and the absolute states, respectively (cf. Section 3 above). Since English does not have a form which is equivalent to the construct state, we expect the distribution of constructions involving the construct state (e.g., possessive noun-noun constructions) to differ between O_{heb} and T_{en}.

POSSESSIVE This feature set contains only one feature indicating whether a possessive suffix is attached to the token. Since Hebrew has several ways of expressing possessiveness, one of which is by means of attaching possessive suffixes (cf. Section 3), we expect the distribution of these suffixes to be different across O_{heb} and T_{en}. In Figure 1, a possessive suffix is attached to the second example.

PREFIX_1, PREFIX_2 Even though Hebrew words can take several prefixes, it is rarely the case that more than two prefixes are attached to one token. We therefore consider only the first two prefix positions as feature categories. We expect them to convey significant classification cues, since they correspond to *function words*: recall that the definite article, the conjunction *and*, and numerous prepositions are realized as prefixes in Hebrew. In Figure 1, the first example exhibits two prefixes: the value of PREFIX_1 is *conjunction* and the value of PREFIX_2 is *temporalSubConj*.

The values of the morphological and POS feature sets are the frequencies of those features within a chunk. We also experimented with the logarithm of the frequencies as the actual values of features, but this turned out to be beneficial only for two classifiers, namely for the POS and the BINYAN classifiers. We therefore use log frequencies for these two feature sets.

Note that we do not define a feature for *every* coordinate of the morphological analysis provided by the analyzer. For example, we find *gender*, *tense*, and *number* to be less rele-

vant for our task. First, we consider them more lexical than morphological, not least due to Hebrew's grammatical gender. Second, these features typically do not reflect translators' decisions, as they are imposed by the source text (unlike, e.g., possessive or passive constructions, where translators have several alternatives to choose from).

5.3 Features Based on Character n-grams

Following Popescu (2011), we experiment with character n-grams. The feature set he designs contains character 5-grams, irrespective of word boundaries. Unlike him, we experiment with 1-grams through 5-grams, as well as with the union of all of them. Presumably, longer n-grams would capture many lexical phenomena, and would thus yield accurate in-domain but inaccurate out-of-domain classifiers (Hebrew words tend to be rather short due to lack of vowels). We also do not go beyond the word level; that is, we calculate n-grams occurring only within one token,[10] since n-grams spanning over several tokens are expected to capture *syntactic* properties of the language, whereas the focus of this study is on morphological features.

Inspired by Koppel et al. (2006), who use Hebrew and Aramaic prefixes and suffixes as features for the classification of rabbinic manuscripts, we design another feature set; we collect bigrams occurring at word boundaries, i.e, at the beginning and the end of tokens.[11] Unlike them, we do not employ a predefined list of suffixes and prefixes. Note that since each bigram in this feature set is preceded or followed by a reserved character marking a word boundary (see Footnote 10), the bigrams in this experiment are, in point of fact, trigrams (in other words, this feature set is a proper subset of the character trigram feature set).

5.4 Features that Approximate Word Structure

We also design a set of features that reflect, on the one hand, the formal representation of morphological information (i.e., the way morphological features are expressed in the or-

[10] Word boundaries are counted as characters. So, for example, the bigrams corresponding to a word like *ab* are *{_a, ab, b_}*, where '_' is a reserved character marking a word boundary.

[11] Volansky et al. (forthcoming) apply a similar feature set to the identification of English translationese.

thography), but, on the other hand, are as content- and domain-independent as possible – that is, features that do not directly reflect lexical information. To this end, we define an abstraction mechanism which is expected to approximate Hebrew word structures, e.g., verbal and nominal patterns.

The idea is to reduce the Hebrew alphabet to a smaller alphabet, allowing symbols in the reduced alphabet to capture sets of characters. We run experiments with three such abstract alphabets (AbA), listed here in decreasing order of abstraction:

AbA$_1$ Consists of only two symbols: *C*, representing all consonants and *V*, replacing the characters traditionally known as *matres lectionis* (cf. Section 3). These characters play a significant role in Hebrew derivational morphology, among other things representing some of the vowels. Formally: AbA$_1$:= $\{C, V\}$, where *C* represents the consonants *b, g, d, z, x, T, k, l, m, n, s, y, p, c, q, r, $, t* and *V* represents *a, h, w, i*.

AbA$_2$ In this alphabet, *C* is as above, but *V* is spelled out. AbA$_2$ thus contains five symbols: $\{C, a, h, w, i\}$. Not only do the *matres lectionis* play a significant role in nonconcatenative morphology (e.g., in verbal and nominal patterns), but the prefixes *h* and *w* also reflect the definite article and the coordinating conjunction *and*, respectively.

AbA$_3$ Includes ten symbols: $\{C, a, h, w, i, b, k, l, m, t\}$, where *C* stands for all remaining letters. The spelled out consonants *b, k, l*, and *m* are prepositions which are realized as prefixes. The characters *k* and *m* can also reflect other grammatical properties, such as gender, number, possessiveness, and tense. The consonant *t* participates in the construction of many verbal and nominal patterns; e.g., it is part of the unmarked feminine plural suffix *wt*.

Figure 2 illustrates how the surface token *wk$hmcxiqwt* ("and when the funny ones [feminine]...") is represented in each of the three abstract alphabets. The feature values in the AbA experiments are (the frequencies of) complete abstracted tokens.

Since no language-specific processing tools are necessary in order to create these ab-

Surface	w	k	$	h	m	c	x	i	q	w	t
AbA_1	V	C	C	V	C	C	C	V	C	V	C
AbA_2	w	C	C	h	C	C	C	i	C	w	C
AbA_3	w	k	C	h	m	C	C	i	C	w	t

Figure 2: The three different abstract representations of the surface form *wk$hmcxiqwt*.

stract representations, applying them to other languages with nonconcatenative morphology (specifically Arabic) is straightforward.

5.5 Feature Combinations

We define two ways of combining features: *disjunction* and *conjunction*. The **disjunction** $F_1 \cup F_2$ results in the union of the feature sets F_1 and F_2. Although the feature vector grows as a result of the disjunction, the features and their values remain the same. Combining by means of disjunction allows for a better understanding of the contribution each feature subset makes.

The **conjunction** $F_1 \times F_2$, on the other hand, results in a new feature set altogether, namely, the Cartesian product of F_1 and F_2. In this study, we employ conjunction in order to enrich the different AbA and character n-gram feature sets with POS information. For example, consider the conjunction of character bigrams and POS; given the input word *hlk* ("walked"), which is tagged as a verb, the features extracted from it are the pairs $\langle$*_h*, verb$\rangle$, $\langle$*hl*, verb$\rangle$, $\langle$*lk*, verb$\rangle$, and $\langle$*k_*, verb$\rangle$; the feature space includes the Cartesian product of all possible bigrams with all 25 POS tags. Similarly, when combining $AbA_1 \times$ POS, each AbA_1 feature, e.g., *CVC*, results in a feature set of 25 features, one for each POS tag: $\langle$*CVC*, noun$\rangle$, $\langle$*CVC*, verb$\rangle$, etc. By applying POS conjunction to the AbA alphabets and character n-grams, we obtain more nuanced and better interpretable feature sets, which remain, nevertheless, abstract and content-independent.

6 Results

We implemented the features discussed in the previous section and constructed SVM classifiers based on each set of features. We then tested the accuracy of each of the classifiers in four experimental setups corresponding to the corpora introduced in Section 4:

InC (*in-corpus*) The training data, i.e., O_{heb} and T_{en}. It includes 600 chunks: 300 original Hebrew instances and 300 translated from English; evaluation is done using ten-fold cross-validation.

InD$_{fr}$ Testing on the *in-domain* dataset containing 30 chunks of twentieth-century literature translated into Hebrew from French. Note that this dataset contains only chunks translated from French, that is, no texts originally written in Hebrew.

OoD-soc Testing on the *out-of-domain* dataset dealing with social science topics (26 chunks, evenly balanced: 13 original Hebrew, 13 translated from English).

OoD-eco Testing on the *out-of-domain* dataset dealing with economics (26 chunks, evenly balanced: 13 original Hebrew, 13 translated from English).

For all the experiments we report accuracy; the baseline (choosing at random) is always 50%, as the test set is balanced. Since the test corpora are relatively small (30 chunks in the case of InD$_{fr}$ and 26 in the out-of-domain experiments), most differences in accuracy on these test sets are not statistically significant. However, differences of a few percentage points on the training set, for which we conduct cross-validation evaluation, are typically significant. To emphasize that, we graphically depict 95% confidence intervals (Clopper & Pearson, 1934) for the results of some of the InC experiments. Complete ranked confidence interval plots for all the experiments (InC, InD$_{fr}$, OoD-soc, and OoD-eco) are listed in the appendix.

6.1 Classifiers Based on Tokens

The accuracies of the classifiers trained on token-based features are given in Table 2. As we conjectured, the classifier trained on word unigrams is highly accurate in the in-corpus

scenario, but does not scale up to the in-domain and out-of-domain datasets. Similarly, and like Koppel and Ordan (2011), we find that a classifier trained solely on function words, while achieving convincing in-corpus results, does not perform very well when applied in other experimental scenarios.

Classifier	InC	InD_{fr}	OoD-soc	OoD-eco
Word unigrams	98.3	66.6	65.3	69.2
Function words	93.5	66.6	69.2	61.5

Table 2: Results of classifiers based on tokens.

6.2 Classifiers Based on Morphological Analysis

The results of the classifiers that reflect morphological aspects, namely, the ones trained solely on the output of the morphological analyzer, are given in Table 3. The confidence intervals of the cross-validation experiments are plotted in Figure 3.

Classifier	InC	InD_{fr}	OoD-soc	OoD-eco
POS	76.8	80.0	50.0	50.0
BINYAN (BI)	57.0	66.6	57.6	53.8
STATUS (ST)	60.2	70.0	53.8	38.4
POSSESSIVE (PS)	54.2	50.0	42.3	46.1
PREFIX_1 (P1)	66.5	60.0	61.5	57.6
PREFIX_2 (P2)	56.3	50.0	46.1	65.3
POS ∪ BI	77.0	83.3	46.1	53.8
POS ∪ ST	77.5	83.3	53.8	50.0
POS ∪ PS	76.5	83.3	53.8	53.8
POS ∪ P1	80.8	76.6	57.6	57.6
POS ∪ P2	78.0	66.6	50.0	61.5
POS ∪ BI ∪ ST ∪ PS ∪ P1 ∪ P2	82.7	66.6	53.8	65.3
BI ∪ ST ∪ PS ∪ P1 ∪ P2	71.5	56.6	69.2	65.3

Table 3: Results of classifiers based on morphological analysis.

The POS classifier, here used mainly as a baseline to test the contribution of the 'pure' morphological features, yields 76.8% accuracy in the in-corpus scenario. While performing quite well on the separate dataset of literary texts translated from French (80%), it fails to

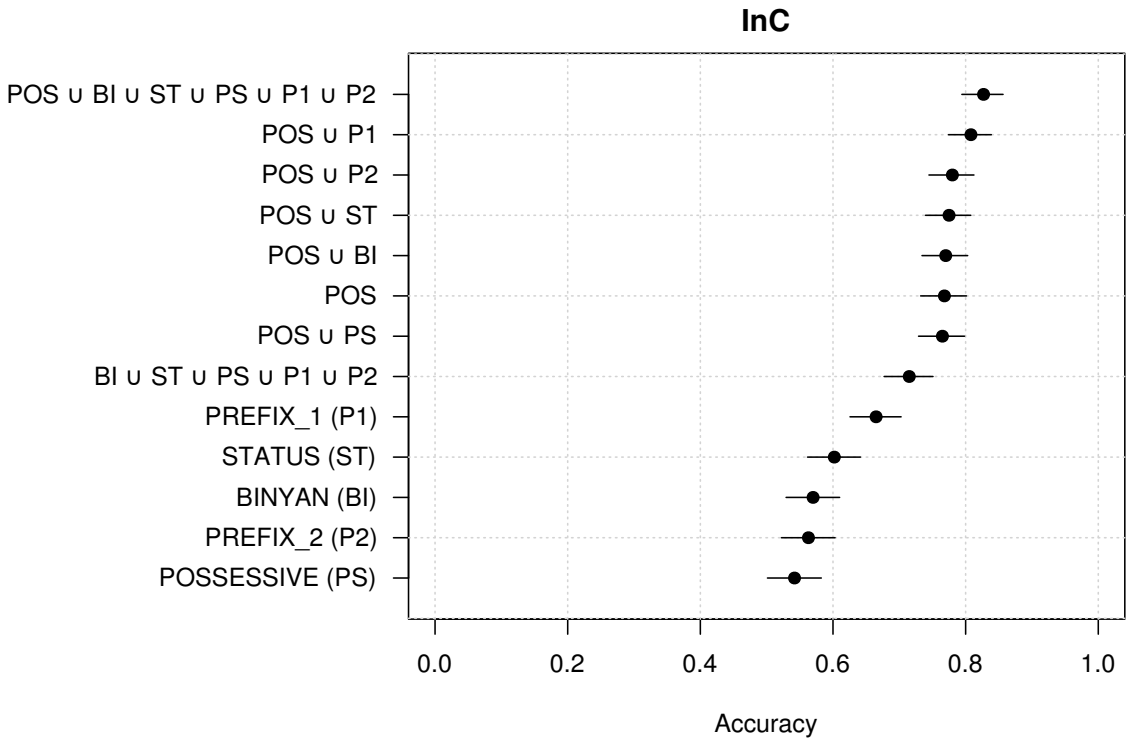

Figure 3: Confidence intervals, classifiers based on morphological analysis (InC scenario).

identify translationese in the OoD datasets. Interestingly, Baroni and Bernardini (2006) report that a similar classifier obtains 49.6% accuracy on their Italian data. They note that "the strikingly low performance of the unigram [POS] model is not surprising, since this model is using the relative frequency of 50 [POS] tags as its only cue" (p. 267). Volansky et al. (forthcoming), on the other hand, obtain 90% when applying a similar feature set to identifying English translationese.

Apart from PREFIX_1, and the somewhat less impressive BINYAN, no other classifier based on a single morphological feature (including POS) manages to perform better than the baseline in all four experimental scenarios. When the features are combined by means of disjunction, the results improve somewhat. Not surprisingly, the combined feature sets that yield the best results in the in-corpus scenario are the ones that uses both POS and PREFIX_1 (POS ∪ P1; POS ∪ BI ∪ ST ∪ PS ∪ P1 ∪ P2). Once the POS baseline is removed and a classifier is trained using only the combination of the single, pure morphological features (BI ∪ ST ∪ PS ∪ P1 ∪ P2), accuracy drops in all scenarios except OoD-soc. Interestingly, this

OoD-soc results is the best any of the pure morphological classifier yields.

In sum, classifiers based on features produced by morphological analysis fail to produce accurate classification results, especially out of domain. The reason may be the low quality of the morphological processing tools we use or the low dimensionality of these classifiers (sometimes containing only one or two features), coupled with the relatively small size of the training set.

6.3 Classifiers Based on Character n-grams

Capturing much lexical information, classifiers based on character n-grams unsurprisingly yield good results (cf. Popescu, 2011; Volansky et al., forthcoming). These results are given in Table 4, and the confidence intervals of the InC experiments are plotted in Figure 4.

Classifier	InC	InD_{fr}	OoD-soc	OoD-eco
1-grams	69.2	26.7	61.5	61.5
2-grams	88.5	66.7	69.2	61.5
2-grams at word boundaries	92.5	70.0	57.6	53.8
3-grams	98.5	66.7	69.2	76.9
4-grams	98.8	73.3	73.1	80.7
4-grams, top-60 features	93.5	66.7	80.7	80.7
5-grams	98.3	66.7	73.1	76.9
1- ∪ 2- ∪ 3- ∪ 4- ∪ 5-grams	98.8	73.3	76.9	69.2
2-grams × POS	98.3	73.3	65.3	57.6
2-grams at word boundaries × POS	97.3	66.7	73.1	80.7
3-grams × POS	98.7	73.3	80.7	61.5
4-grams × POS	98.5	73.3	80.7	69.2

Table 4: Results of classifiers based on character n-grams.

The optimal n for n-gram classifiers seems to be 4; not only is the 4-gram classifier highly accurate in cross-validation, it scales up nicely to out-of-domain tasks. Extending n-gram length to 5, or taking all n-grams of lengths 1 to 5, does not seem to improve much. Enhancing the n-grams with POS information brings about a small accuracy gain in most cases.

As a further indication of the robustness of the n-gram classifiers, we experimented with a 4-gram classifier that only uses the 30 most indicative features of O_{heb} and the 30 most

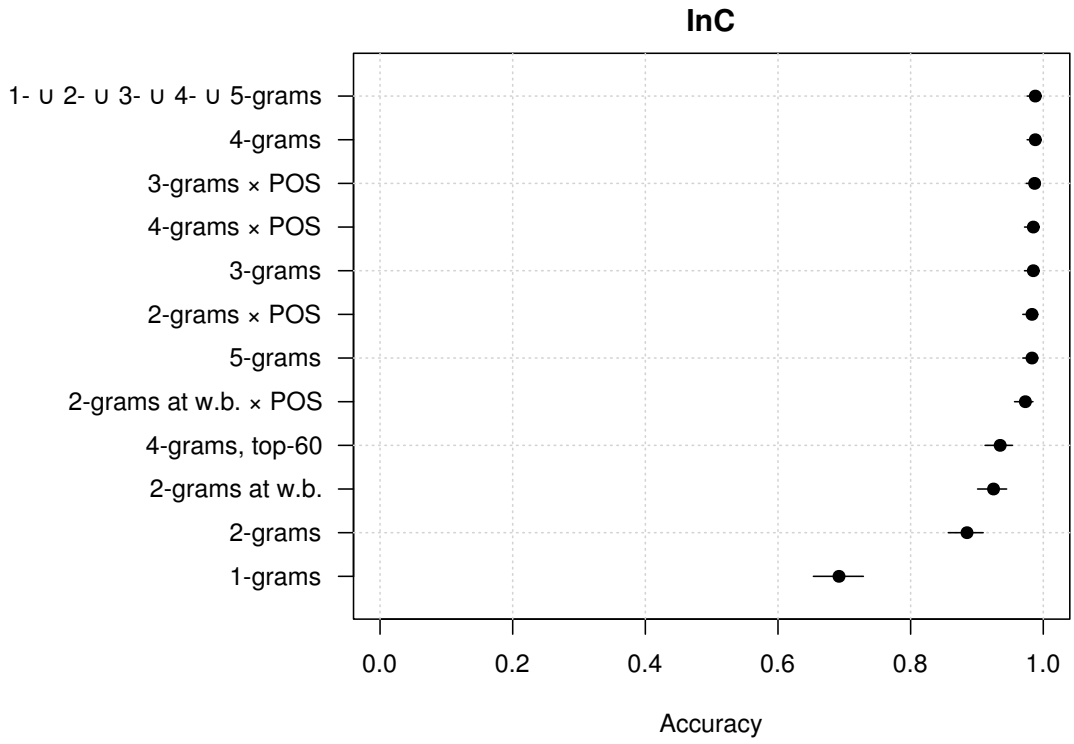

Figure 4: Confidence intervals, classifiers based on character n-grams (InC scenario).

indicative of T_{en}: after training a classifier with the entire set of 4-grams, we selected the 30 features whose weights were greatest (most indicative of O_{heb}) and the 30 features whose weights were lowest (most indicative of T_{en}). We then trained a new classifier using only these 60 features.[12] The results of this top-60 4-gram classifier are listed in Table 4, and demonstrate the power of simple, low-dimensional classifiers. Some of the most indicative features are discussed in Section 7.

6.4 Classifiers Based on Abstract Alphabets

The results of the classifiers that approximate Hebrew word structure by means of alphabet abstractions are given in Table 5, and the confidence intervals of the InC experiments are plotted in Figure 5.

AbA_1 and AbA_2 reveal mixed results. While performing extremely well in certain scenarios (for example, the result AbA_2 obtains on the OoD-eco dataset, 96.1%, is by far the best

[12] The specific features are listed in the appendix.

Classifier	InC	InD_{fr}	OoD-soc	OoD-eco
AbA_1	78.2	46.6	35.6	76.9
AbA_2	92.2	43.3	73.1	96.1
AbA_3	97.0	70.0	80.7	65.3
AbA_1 × POS	92.0	63.3	65.3	73.1
AbA_2 × POS	95.8	76.6	73.1	80.7
AbA_3 × POS	98.0	86.6	84.6	65.3
AbA_3 × POS, top-40 features	91.7	80.0	80.7	76.6

Table 5: Results of classifiers based on abstract alphabets.

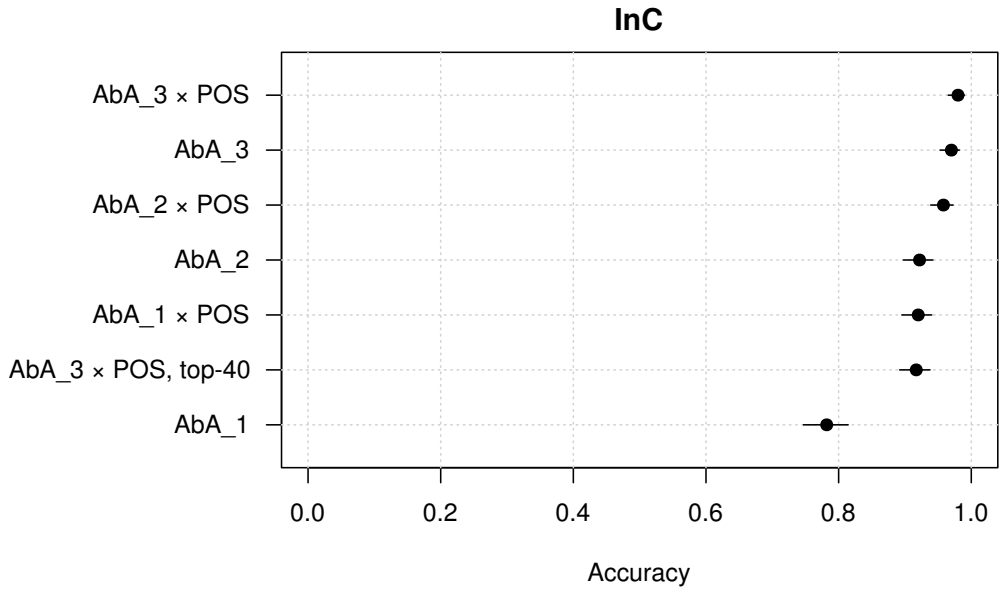

Figure 5: Confidence intervals, classifiers based on abstract alphabets (InC scenario).

one any of our classifiers yields in that scenario, cf. Figure 9 in the appendix), they fail miserably in other scenarios. In contrast, AbA_3 yields competitive results in all four scenarios. Considering the fact that these results are obtained without applying any feature selection methods, they are very promising. It stands to reason that by cautiously reducing the feature spaces of these simple AbA classifiers, their performance will increase significantly. We intend to do that in future work.

Upon enriching the abstract templates with POS information by means of conjunction, we observe accuracy improvement in most scenarios. While all three classifiers yield results well above the baseline, $\text{AbA}_3\times$ POS is the best overall classifier we train in this study. Here,

too, we experimented with a classifier using only the top-n features most indicative of O and the top-n most indicative of T, this time with n being 20. This very low-dimensional classifier is the only one in this study obtaining more than 75% in *all* experimental scenarios.

7 Analysis

We now look more closely at the results discussed in the previous section. Specifically, in order to understand which features are more relevant than others for a given classifier, we examine the weights an SVM assigns to the features it uses: the higher the weight, the more important the feature is considered to be. Note that low weights are not always an indication that a feature is not important, as potential dependencies among features can discount important features. The inverse, however, does not hold: a feature assigned a high weight indicates a significant property of one of the classes to be distinguished. In the following, we highlight some of the more successful discriminating features.

Word unigrams The token *dwwqa* is one the most prominent markers of O_{heb}, i.e., it is underrepresented in translated texts. *dwwqa* is an adverb that roughly means "contrary to expectations." Importantly, it is not lexicalized in English. This is a typical case of *negative interference* (Toury, 1995): a lexical gap between the two language systems involved in the translation process creates a situation where nothing in the source language (in this case, English) triggers the generation of the lacking item (in this case, *dwwqa*) when translating into the target language (Hebrew). In our corpus, *dwwqa* is almost six times more frequent in O_{heb} than in T_{en}.

POS Three major differences in the POS distribution between T_{en} and O_{heb} are given in Table 6.

The difference in the distribution of proper names goes hand in hand with the *explicitation hypothesis* (Blum-Kulka, 1986). According to this hypothesis, translators tend to render implicit utterances in the source text (e.g., pronouns) more explicitly in the target text they

POS tag	O_{heb}	T_{en}	Ratio (T_{en}/O_{heb})
properName	22,808	28,470	1.25
copula	9,463	11,403	1.21
modal	3,243	3,677	1.13

Table 6: Three major differences in the POS distribution across T_{en} and O_{heb}.

produce, specifically by means of proper names. We note, however, that we (unlike, e.g., Baroni & Bernardini, 2006) do not notice remarkable differences in the distribution of pronouns between O_{heb} and T_{en}.

We also find that modal constructions are more frequent in translated than in non-translated Hebrew texts. This is a case of *positive interference* (Toury, 1995), i.e., overrepresentation of features characteristic to the source language in translations (in this case, from English). Interestingly, the different distribution of modal verbs across T_{en} and O_{heb} provides us with a partial explanation of the excessive use of copulas in translated texts: in Hebrew, most modal verbs do not inflect for tense; in order to express the past tense, modals are combined with copula past tense forms (functioning in these constructions as an auxiliary verb). And indeed, we find in our data that copulas in the past tense are highly collocated with modal verbs. This partially explains why copulas are more frequent in T_{en}. Another contributing factor stems from the optionality of Hebrew present-tense copulas in non-verbal sentences (Haugereid, Melnik, & Wintner, 2013); since counterpart copulas are mandatory in English, they tend to be explicated in translations to Hebrew.

Morphological features The classifiers trained on the pure, single morphological feature sets do not perform very well, neither in the ten-fold cross-validation in-corpus scenario, nor when tested on the additional in-domain and out-of-domain datasets. This might be due to the low dimensionality of these classifiers, to the relatively small amount of training data, or to the performance of the morphological analyzer, which is often inaccurate.

Perhaps not surprisingly, the classifier based on PREFIX_1 is the best performing one, as it reflects linguistic information which is realized in other languages (like English) as func-

tion words, e.g., conjunctions and prepositions. We find that the most significant marker of O_{heb} is the prefix corresponding to the coordinating conjunction *and*. Indeed, this prefix is 1.24 times more frequent in O_{heb} than in T_{en}. This finding calls for further research by translation scholars.

Even though the BINYAN classifier, based on the seven Hebrew verbal patterns, manages to perform slightly better than the baseline in all experimental scenarios, we cannot interpret the results it yields. The reason is that the accuracy of the morphological processor is particularly poor with respect to the verbal patterns.

Character *n*-grams Many of the most discriminative word unigrams are also reflected in the results of other feature classes like character *n*-grams and abstract alphabets. So, for instance, the character trigrams *_dw*, *dww*, *wwq*, *wqa*, and *qa_* (corresponding to the token *dwwqa* mentioned above) are amongst the most prominent markers of O_{heb}. In other words, even though we set out to capture morphological properties by looking at sub-tokens and alphabet abstractions, we sometimes end up capturing lexical cues.

A detailed analysis of the most significant features of the 4-gram classifier reveals the following pattern: among the ten strongest indications of original Hebrew are three substrings of the lexical *dwwqa*, followed by *ywd* "more/still/yet", *kbr* "already", *gm* "also" and *mwl* "against", with indications of word boundaries at either side of these short words. Note that these are all function words, that likely do not have direct, one-to-one counterparts in the source languages, and hence are distributed very differently between O_{heb} and T_{en}. Other indications of O_{heb} include *bi$r* (the prefix of *bi$ral* "in Israel") and *_tl_* (with word boundaries at both ends), clearly referring to Tel Aviv.

Strong indications of translations include the prepositions *kdi* "in-order-to" and *bzmn* "while/during", the adjective *nwsp* "additional", the modal *yewi* "may", but also *n*-grams that are more abstract and less transparent.

AbA$_1$ We find that one of the most prominent features of T_{en} is a triplet of *matres lectionis*, namely *VVV*. This template mostly encompasses four tokens which play a crucial role in

Hebrew grammar: 1. *hwa*, which can be either a pronoun ("he") or a third person singular copula in the present tense ("is") 2. *hia* ("she"), which is the same as *hwa*, namely, both a pronoun and a copula, only feminine 3. *hih* ("was") and 4. *hiw* ("were"), which are copula forms in the past tense. Table 7 illustrates how these four most characteristic instantiations of this template are realized across O_{heb} and T_{en}. Together they constitute 96% of this template's occurrences in the training data. A reason for the overrepresentation of copula in T_{en}, namely, *positive interference*, was discussed above.

VVV_{AbA_1}	O_{heb}	T_{en}	Ratio (T_{en}/O_{heb})
hwa	4,227	5,617	1.33
hia	3,431	3,913	1.14
hih	2,835	3,607	1.27
hiw	908	1,394	1.54

Table 7: The four most characteristic instantiations of the AbA_1 template *VVV* and their distribution in T_{en} and O_{heb}.

By looking at an abstract feature as simple as the *VVV* sequence, which potentially leaves room for 4^3 surface forms (in practice, only 30 of them are realized in the training data), we already have at our disposal numerous highly frequent distinguishing markers.

AbA_2 Naturally, some of the results found in AbA_1 are also reproduced in AbA_2. For example, the spelled-out instance *hwa* of the AbA_1 *VVV*-template is one of the top markers of T_{en}.

Another discriminating marker of T_{en} captured by AbA_2 is the template *hCCia*, which captures certain instances of the verb pattern *Hif'il* (in past tense, third person singular masculine), namely, those instances with *a* as the third (and final) letter of the root. The *Hif'il* pattern is predominantly used as a causative in Hebrew and might indicate a structural difference between English and Hebrew. This calls for further investigation.

The same AbA_2 template, *hCCia*, also captures the token *hn$ia* ("the president"), reflecting perhaps a cultural marker (the Israeli head of state is the prime minister, rather than the president). Although lexical, this feature captures a significant difference between T_{en}

and O_{heb} which is scalable to other domains, since it is rather frequent in genres such as newspaper articles and parliament proceedings.

AbA$_3$ While less abstract than AbA$_1$ and AbA$_2$, this third alphabet touches on morphological templates that cannot be captured with the more abstract alphabets. Consider the template *mCwCl*, which, theoretically speaking, exclusively captures the masculine singular passive participle of roots whose third (and final) letter is *l*. This template is three times more frequent in T_{en}. The most frequent instance of this template, the modal *mswgl* ("capable of"), reflects constructions which are, as discussed above, more frequent in translated texts due to *positive interference*. The other instances of this AbA$_3$ template suggest that there are different distributions of morphological items between T_{en} and O_{heb}. This, too, calls for further studies.

AbA, general Importantly, we manage to capture with the AbA templates many discriminative markers – whether lexical, morphological, or (morpho)syntactic – without relying on a ready-made, morphologically or syntactically informed mechanism.

Enriching the templates with POS information improves the results. Once the AbA templates are restricted to capture smaller token spaces, they become more precise. In AbA$_1$ × POS, a prominent AbA$_1$ feature like *VVV* is spelled out into 25 features (each corresponding to a different POS tag), thereby making ⟨*VVV*, pronoun⟩ and ⟨*VVV*, copula⟩ much more dominant than, say, ⟨*VVV*, noun⟩. The POS enhancement thus helps the classifiers to separate the wheat from the chaff.

8 Conclusion

We have employed text classification for the investigation of translationese in a morphologically complex language, namely Modern Hebrew. This is the first work addressing the automatic identification of translationese in a Semitic language, and the first focusing on the morphological manifestation of translated texts' properties. Specifically, we have trained

several SVM classifiers that distinguish with high accuracy between twentieth-century literary texts translated from English and similar texts originally written in Hebrew. Some of these classifiers have proven to be robust, yielding good results when tested on different datasets, i.e, on texts from the same domain (twentieth-century literature), but translated from a different source language (French), and on texts from other domains, namely newspaper and journal articles dealing with the social sciences and economics. The fact that some of the classifiers scale up to other experimental scenarios supports our hypothesis that training on a corpus of contemporary literature – a very heterogeneous dataset – is suitable and beneficial for the development of scalable classifiers.

Numerous feature design strategies have been explored: function words, word unigrams, pure morphological features, POS tags, character n-grams, and three different instances of a novel alphabet abstraction mechanism aimed at approximating Hebrew word structure. We have also experimented with several hybrid feature sets, i.e., combinations of some of the aforementioned feature sets by means of disjunction and conjunction.

Classifiers trained solely on morphological information do not obtain very good results; this might be due to the performance of the often inaccurate morphological analyzer, the low dimensionality of these classifiers, or the relatively small amount of training data. The classifiers obtaining the best overall results use combined models, conjunctions of POS information with either an alphabet abstraction or character n-grams. This indicates that, currently, the best way to represent word-level and sub-word-level phenomena in Hebrew, for the purpose of identifying translationese, is by approximating morphological analysis using shallow abstractions, and restricting those abstractions to specific POS spaces.

As we saw in the previous section, even though we set out to capture morphological properties of Hebrew, we sometimes end up capturing non-morphological features. Even when applying POS annotation and alphabet abstractions, lexical markers manage to "sneak in," e.g., in the form of proper nouns. Indeed, some of the most significant classification cues do not reflect morphological traits of the Hebrew language.

Let us revisit the example of the lemma *dwwqa* (roughly meaning "contrary to expec-

tations"). It appears 6.62 times more often in O_{heb} than in T_{en} (331 vs. 50 occurrences), thus averaging slightly more than one occurrence per original Hebrew chunk; its probability to appear in a T_{en} chunk, on the other hand, is 1/6 (assuming a uniform distribution). Although not a morphological feature, *dwwqa* does reflect a structural difference between Hebrew and English (and French), and due to its relatively high frequency it contributes immensely to classification. For example, *CVVCV* is one of the most significant features in the AbA_1 experiment. Similarly, as described above in Section 7, n-grams corresponding to substrings of *dwwqa* are amongst the most prominent markers of O_{heb}. In this sense we conclude that although our abstractions are not purely morphology based, the non-morphological features that do manage to sneak in are of both theoretical and practical value.

In future work, our first concern will be dimensionality reduction. As we show with the 4-gram and AbA_3 × POS classifiers, a very low-dimensional space of only 40 or 60 features suffices for producing highly accurate results. We intend to employ state-of-the-art algorithms in order to rank feature sets and select the most discriminative feature subsets, thereby reducing the size of feature vectors and limiting the effect of overfitting to the training data. This should bring about accuracy gains, but also facilitate a better understanding of the morphological properties of Hebrew translationese.

We also intend to explore other ways of designing alphabet abstractions, more sophisticated than the ones developed and discussed in the present work. Substitutions could, for example, be made dependent upon positions within the surface token, thereby simulating Hebrew prefixes and suffixes. In this study, we have combined abstract alphabets only with POS tags. This has turned out to be a promising approach. Combinations with other feature sets might also prove fruitful, in particular with PREFIX_1, the pure morphological feature set yielding the best results. Finally, we plan to apply similar alphabet abstractions to other Semitic languages, such as Arabic, building on the similar root-and-pattern morphological structures of words in these languages.

Acknowledgments

This research was supported by a grant from the Israeli Ministry of Science and Technology. We are grateful to Ted Briscoe for suggesting some of the *n*-gram experiments and for useful discussions. We wish to thank Titus von der Malsburg for suggesting the use of confidence intervals. We are also grateful to Bracha Lang for providing us with the out-of-domain corpora, and to Kayla Jacobs for providing us with the list of Hebrew function words. Thanks are also due to Irit Noy for annotating the literary texts with additional metadata.

References

Argamon, S. (2011). Book review of *Scalability Issues in Authorship Attribution*, by Kim Luyckx. *Literary and Linguistic Computing*, *27*(1), 95–97.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). Amsterdam: John Benjamins.

Bar-Haim, R., Sima'an, K., & Winter, Y. (2008). Part-of-speech tagging of Modern Hebrew text. *Natural Language Engineering*, *14*(2), 223–251.

Baroni, M., & Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, *21*(3), 259–274.

Blum-Kulka, S., & Levenston, E. A. (1983). Universals of lexical simplification. In C. Færch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 119–139). Longman.

Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In J. House & S. Blum-Kulka (Eds.), *Interlingual and intercultural communication discourse and cognition in translation and second language acquisition studies* (Vol. 35, pp. 17–35). Tübingen: Gunter Narr.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*(4), 404-413.

Daniels, P. T. (1997). Scripts of Semitic languages. In R. Hetzron (Ed.), *The Semitic languages* (pp. 16–45). Routledge.

Diederich, J., Kindermann, J., Leopold, E., & Paass, G. (2003). Authorship attribution with support vector machines. *Applied intelligence*, *19*(1-2), 109–123.

Fabri, R., Gasser, M., Habash, N., Kiraz, G., & Wintner, S. (2014). Linguistic introduction: The orthography, morphology and syntax of Semitic languages. In I. Zitouni (Ed.), *Semitic language processing* (pp. 3–41). Berlin and Heidelberg: Springer.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In L. Wollin & H. Lindquist (Eds.), (pp. 88–95). Lund: CWK Gleerup.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Haugereid, P., Melnik, N., & Wintner, S. (2013). Nonverbal predicates in Modern Hebrew. In S. Müller (Ed.), *Proceedings of the 20th international conference on head-driven phrase structure grammar* (pp. 6–26). CSLI Publications. Retrieved from `http://cslipublications.stanford.edu/HPSG/2013/hmw.pdf`

Ilisei, I., & Inkpen, D. (2011). Translationese traits in Romanian newspapers: A machine learning approach. *International Journal of Computational Linguistics and Applications*, *2*.

Ilisei, I., Inkpen, D., Pastor, G. C., & Mitkov, R. (2010). Identification of translationese: A machine learning approach. In A. F. Gelbukh (Ed.), *Proceedings of CICLing-2010: 11th international conference on computational linguistics and intelligent text processing* (Vol. 6008, pp. 503–511). Springer.

Itai, A., & Wintner, S. (2008). Language resources for Hebrew. *Language Resources and Evaluation*, *42*(1), 75–98.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, *1*(3), 233–334.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit* (Vol. 5).

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, *17*(4), 401–412.

Koppel, M., Mughaz, D., & Akiva, N. (2006). New methods for attribution of rabbinic literature. *Hebrew Linguistics: A Journal for Hebrew Descriptive, Computational and Applied Linguistics*, *57*, 5–18.

Koppel, M., & Ordan, N. (2011). Translationese and its dialects. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 1318–1326). Portland, Oregon, USA: Association for Computational Linguistics.

Kurokawa, D., Goutte, C., & Isabelle, P. (2009). Automatic detection of translated text and its impact on machine translation. *Proceedings of MT Summit XII*, 81–88.

Lembersky, G., Ordan, N., & Wintner, S. (2011, July). Language models for machine translation: Original vs. translated texts. In *Proceedings of EMNLP*.

Lembersky, G., Ordan, N., & Wintner, S. (2012a, April). Adapting translation models to translationese improves SMT. In *Proceedings of the 13th conference of the european chapter of the association for computational linguistics* (pp. 255–265). Avignon, France: Association for Computational Linguistics. Retrieved from `http://www.aclweb.org/anthology/E12-1026`

Lembersky, G., Ordan, N., & Wintner, S. (2012b, December). Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, *38*(4), 799–825. Retrieved from `http://dx.doi.org/10.1162/COLI_a_00111`

Lembersky, G., Ordan, N., & Wintner, S. (2013, January). Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics*, *39*. Retrieved from `http://dx.doi.org/10.1162/COLI_a_00159`

Lembersky, G., Shacham, D., & Wintner, S. (2014, January). Morphological disambiguation of Hebrew: a case study in classifier combination. *Natural Language Engineering*, *20*, 69–97. Retrieved from `http://journals.cambridge.org/article_S1351324912000216`

Lynch, G., & Vogel, C. (2012). Towards the automatic detection of the source language of a literary translation. In *Proceedings of the 24th international conference on computational linguistics (COLING): Posters* (pp. 775–784).

Mosteller, F., & Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist.* Addison-Wesley.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods: Support vector learning.* Cambridge, MA: MIT Press.

Popescu, M. (2011). Studying translationese at the character level. In G. Angelova, K. Bontcheva, R. Mitkov, & N. Nicolov (Eds.), *Proceedings of recent advances in natural language processing* (pp. 634–639).

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys, 34*(1), 1–47.

Toury, G. (1995). *Descriptive translation studies and beyond.* Amsterdam and Philadelphia: John Benjamins.

Volansky, V., Ordan, N., & Wintner, S. (forthcoming). On the features of translationese. *Literary and Linguistic Computing*.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (Second ed.). Morgan Kaufmann.

Yona, S., & Wintner, S. (2008). A finite-state morphological grammar of Hebrew. *Natural Language Engineering*, *14*(02), 173–190.

Most Distinctive N-gram Features

The 60 most distinctive n-gram features (Section 6.3) are: *wqa_, wwqa, dwwq, ywd_, _kbr, _gm_, mwl_, kbr_, _ywd, _mwl, _dww, bier, bkll, _wrq, _egm, _acl, kll_, ela_, egm_, klwm, _lw_, _klw, wla_, _wla, _tl_, _ph_, _npe, ain_, blnw, sbln, arwx, _yew, _lmy, bmek, ylwl, hwa_, mswg, laxw, _ah_, dmii, _mlb, _zmn, hbiy, _bzm, briq, mlbd, _msw, _ydi, amwr, _keh, lehi,*

*mek_, ydii, mbri, _hwa, yewi, nwsp, bzmn, _kdi, kdi_*

Confidence Interval Plots

We graphically depict below 95% confidence intervals corresponding to the experiments reported on in Section 6: the InC experiments (Figure 6), the InD$_{fr}$ experiments (Figure 7), the OoD-soc experiments (Figure 8), and the OoD-eco experiments (Figure 9).

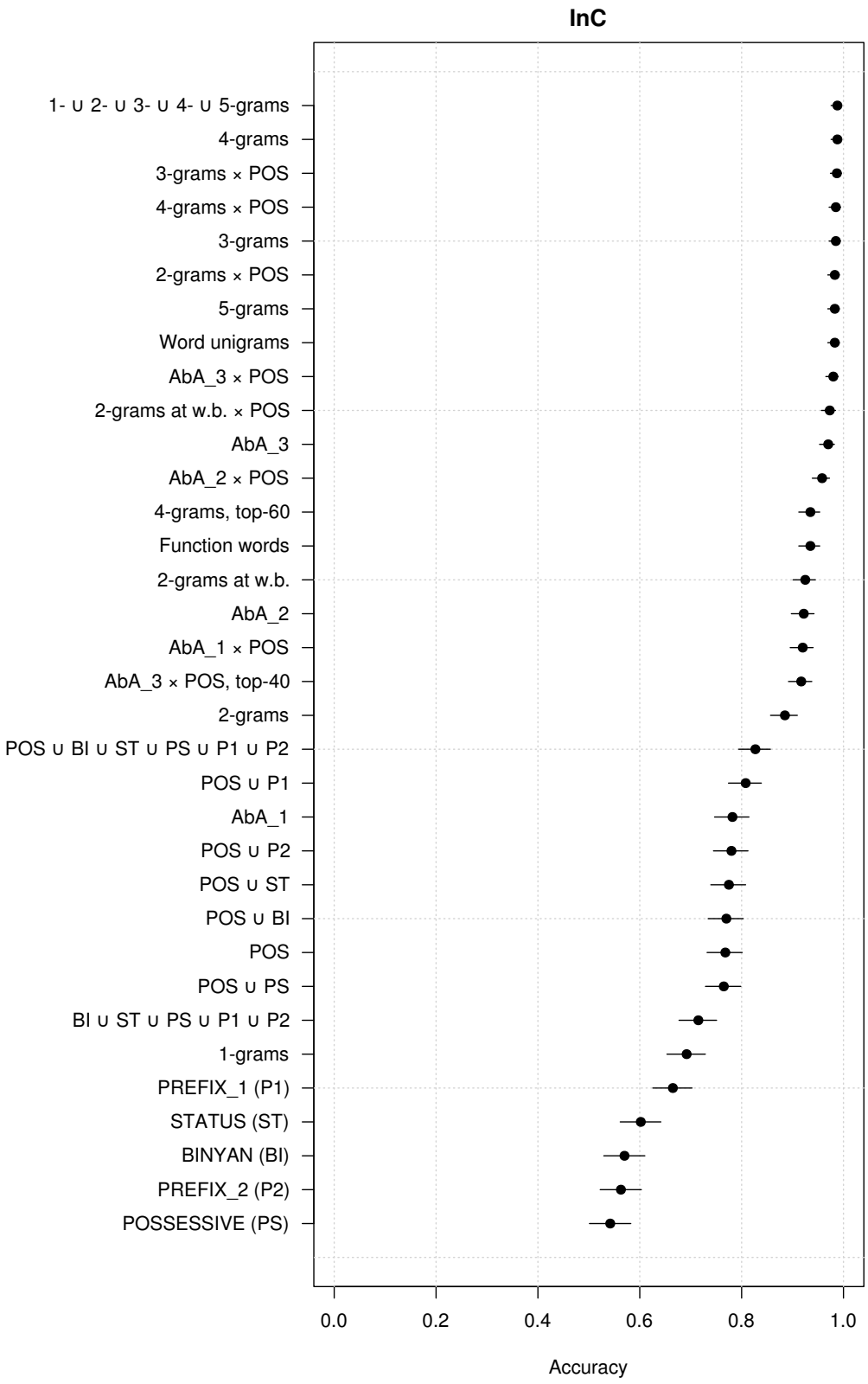

Figure 6: Confidence intervals, InC experiments.

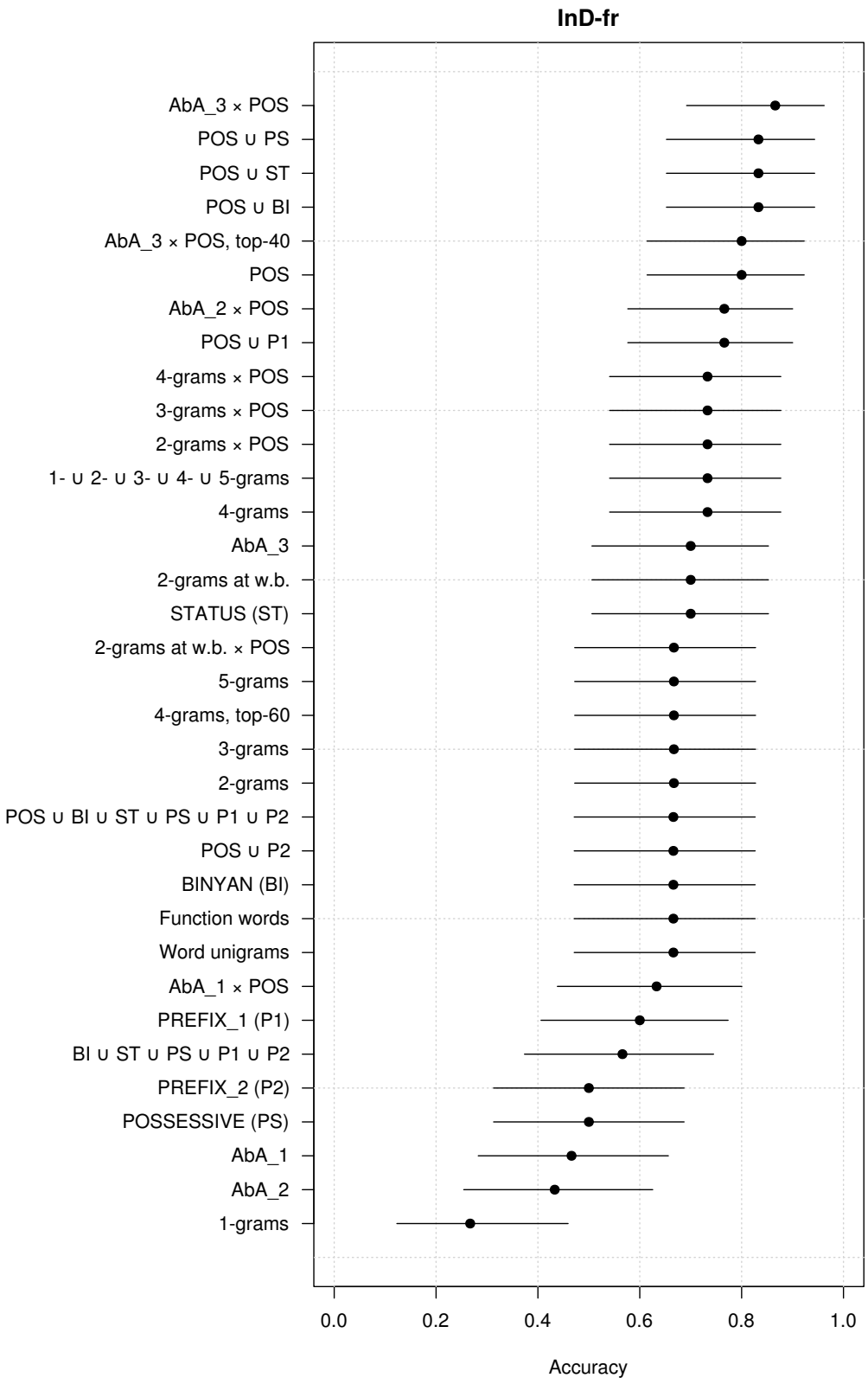

Figure 7: Confidence intervals, InD_{fr} experiments.

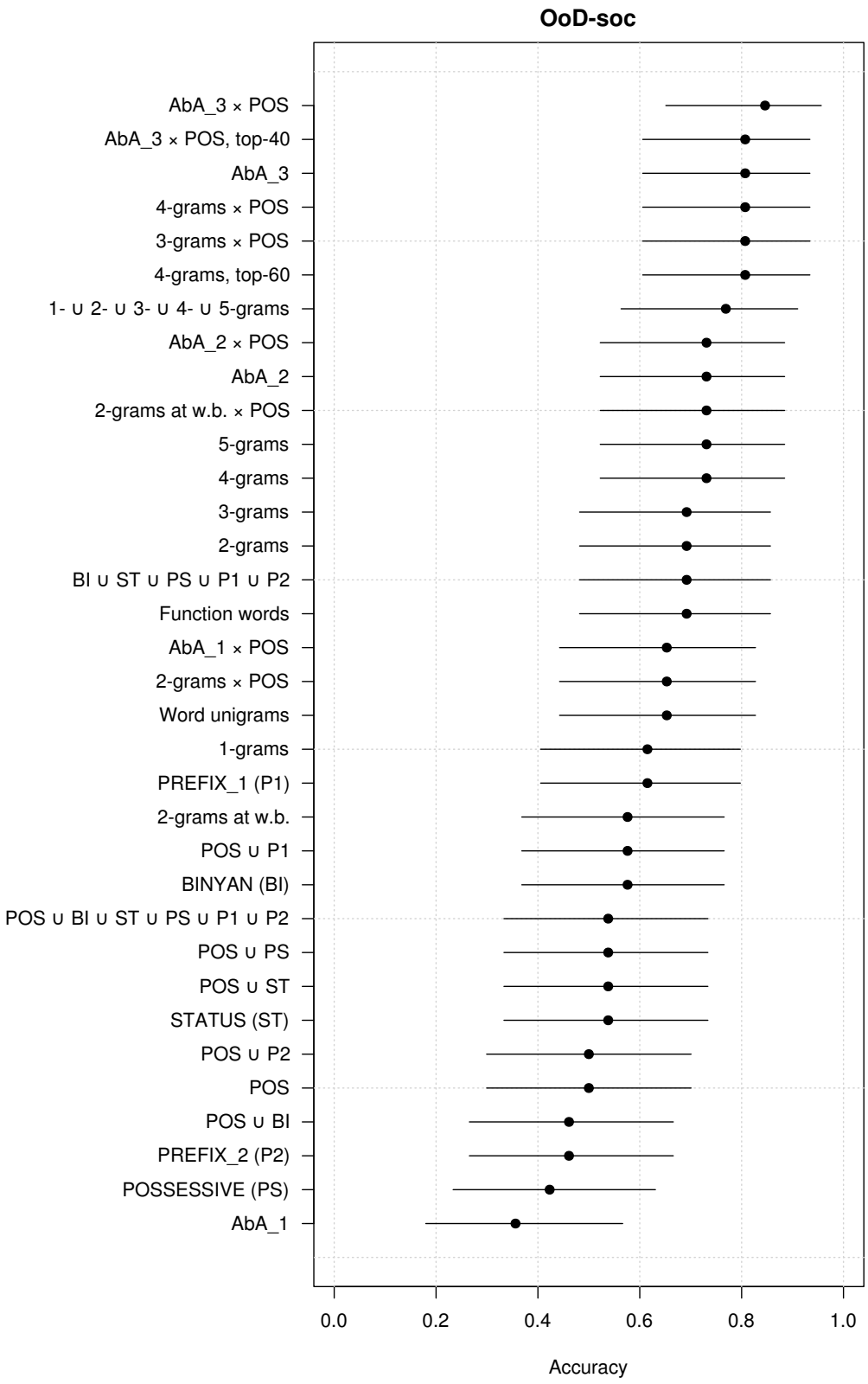

Figure 8: Confidence intervals, OoD-soc experiments.

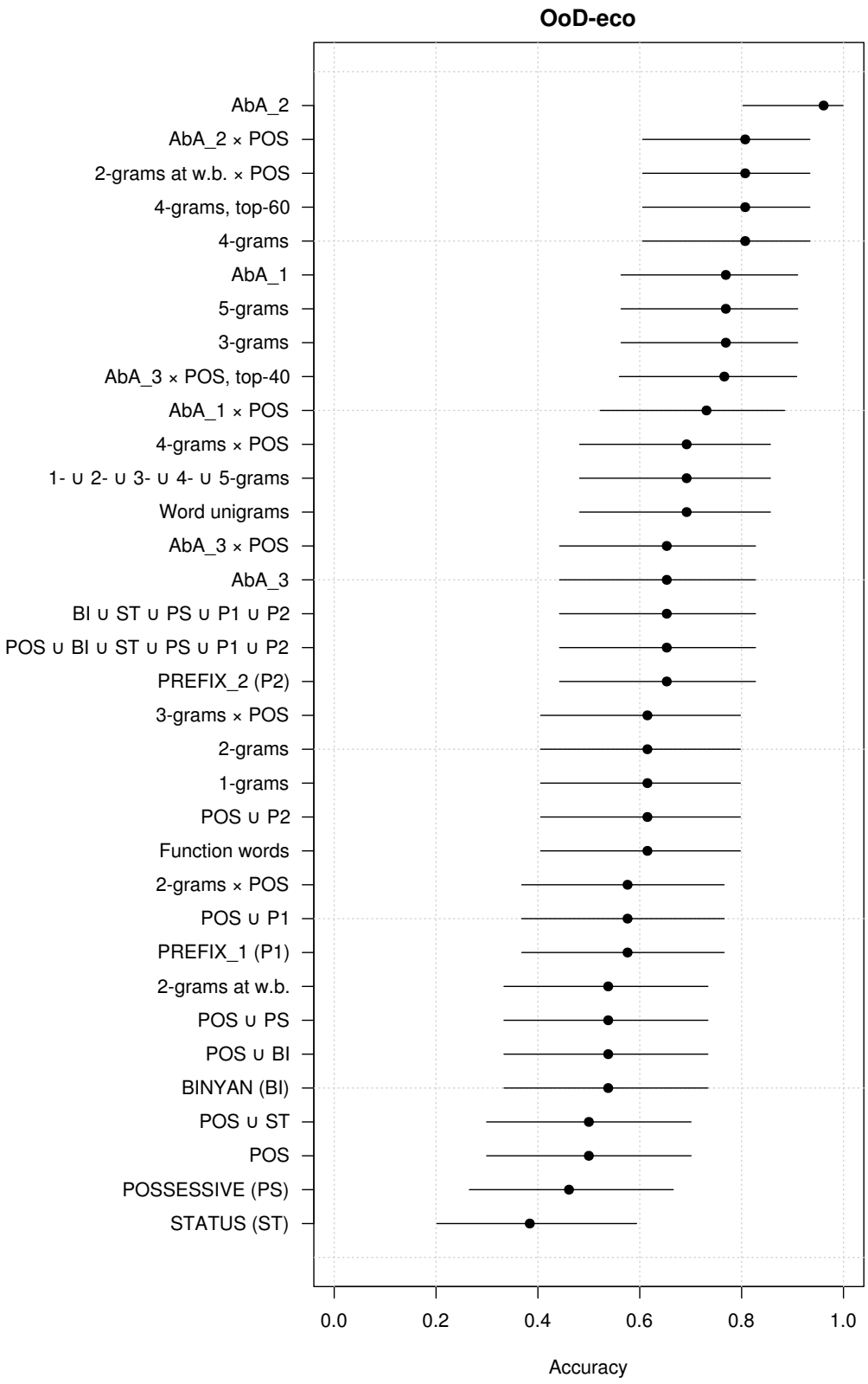

Figure 9: Confidence intervals, OoD-eco experiments.