Frugal Paradigm Completion

Alexander Erdmann,[†] **Tom Kenter, Markus Becker and Christian Schallhart** [*]

Google, UK [†]Ohio State University, USA

`erdmann.6@osu.edu, {tomkenter,mabecker,schallhart}@google.com`

Abstract

Lexica distinguishing all morphologically related forms of each lexeme are crucial to many language technologies, yet building them is expensive. We propose *Frugal Paradigm Completion*, an approach that predicts all related forms in a morphological paradigm from as few manually provided forms as possible. It induces typological information during training which it uses to determine the best sources at test time. We evaluate our language-agnostic approach on 7 diverse languages. Compared to popular alternative approaches, our Frugal Paradigm Completion approach reduces manual labor by 16-63% and is the most robust to typological variation.

1 Introduction

From syntactic parsing (Seeker and Kuhn, 2013) to text-to-speech (Zen et al., 2016; Wan et al., 2019), many linguistic technologies rely on accurate lexica decorated with morphological information. Yet, building such lexica requires much human effort (Buckwalter, 2002; Tadić and Fulgosi, 2003; Forsberg et al., 2006; Sagot, 2010; Eskander et al., 2013). We present a language-agnostic method for minimizing the manual labor required to add new paradigms to an existing lexicon.

Formally, let each lexicon entry, or realization, be a triple (P, C, f). P marks membership in some paradigm $\mathbb{P}$ of morphologically related words, C defines a cell in $\mathbb{P}$ as a bundle of morphosyntactic features, and f is the form realizing C in $\mathbb{P}$. Hence, paradigm $\mathbb{SING}$ can be expressed (in the UniMorph schema (Kirov et al., 2018)) as a set of realizations: {(SING, NFIN, *sing*), (SING, 3.SG.PRES, *sings*), . . . }.

For each paradigm to be added to the lexicon, e.g., $\mathbb{FLY}$, we aim to select as few sources as pos-sible to be manually realized, e.g., {(FLY, NFIN, *fly*), (FLY, PST, *flew*)} such that the forms realizing the remaining cells can be predicted, i.e., *flies*, *flying*, *flown*. Here, sources are manually provided realizations. Targets are realizations whose forms must be predicted from sources. Our work differs from traditional paradigm completion (Durrett and DeNero, 2013) in that sources are not given blindly, but the system must strategically select which sources it wants to be given at test time.

Paradigm completion from one source is typically non-deterministic due to multiple inflection classes realizing different exponents in some cells, e.g., suffixing +*ed* generates the past tense for $\mathbb{WALK}$, but not for $\mathbb{SING}$ or $\mathbb{FLY}$ which are members of different classes. Hence, many works discuss paradigm completion in the context of (implicit) inflection class disambiguation (Ackerman et al., 2009; Montermini and Bonami, 2013; Beniamine et al., 2018). Finkel and Stump (2007) propose three approaches to select the fewest sources required to deterministically identify class. Yet, neural sequence models can often complete paradigms accurately from less sources without fully disambiguating inflection class (Kann and Schütze, 2016; Aharoni and Goldberg, 2017; Wu and Cotterell, 2019). See Elsner et al. (2019) for an overview of the application of neural sequence models to morphological theory.

We propose Frugal Paradigm Completion (FPC), inspired by work on inflection class disambiguation and neural sequence modeling. We train a *source selection agent* (SSA) to induce typological knowledge regarding the distribution of complexity in paradigms and use this to request informative source cells to be realized by an *oracle*. Sources are fed to a *predictor* to generate target forms. For each paradigm, SSA iteratively requests sources until the oracle confirms all cells have been realized correctly.

We introduce a novel metric, *auto-rate*, to quantify the manual labour (performed by the oracle) needed to complete each paradigm. Using this metric, we demonstrate that FPC reduces labor by 63% over predicting targets from lemmata, and 47% over predicting them from the smallest set of sources that fully disambiguates inflection class. We propose a new typology for discussing the organization of complexity in paradigms which helps explain why strategies perform better or worse on certain languages while FPC, being sensitive to typological variation, performs robustly.

After discussing related paradigm completion approaches in Section 2, we describe FPC in Section 3. Section 4 covers all data and experimental set up details. We discuss results in Section 5 and analyze FPC's behavior in Section 6.

2 Paradigm Completion Approaches

Here we discuss several paradigm completion approaches related to FPC.

Lemma-based Paradigm Completion The standard paradigm completion approach does not select sources, but assumes one source: the lemma (Dreyer and Eisner, 2011), whose distinction is ultimately arbitrary. Yet many have shown that more informative sources can be chosen (Finkel and Stump, 2007; Cotterell et al., 2017b; Kann and Schütze, 2018).

Most Informative Source For each target form to be predicted, Kann and Schütze (2018) select the source most likely to predict that form. Unlike FPC, they do not attempt to minimize the number of unique sources that must be manually realized.

Static Principal Parts To minimize sources required to fully disambiguate inflection class, Finkel and Stump (2007); Stump and Finkel (2013) propose three approaches: *static*, *dynamic*, and *adaptive*. In the static approach, the same sources must be used for every paradigm (these sources are referred to as principal parts in a much older pedagogical tradition dating back to ancient Rome with Varro's *de lingua latina* (Grinstead, 1916; Ahern, 1990)). Cotterell et al. (2017b) train a model on static sources and attain near 100% accuracy in Latin verb paradigm completion. However, they do not consider that one paradigm may require fewer sources than another, nor that paradigm completion may require fewer sources than inflection class disambiguation.

Dynamic Principal Parts Finkel and Stump (2007)'s dynamic approach selects a minimal set of sources necessary to fully disambiguate inflection class which can be unique to that inflection class. While efficient, this is impractical in that it requires oracular knowledge of class prior to seeing any forms.

Adaptive Principal Parts Finkel and Stump (2007)'s adaptive approach, like our FPC method, chooses the same first source cell for each paradigm $\mathbb{P}$. Subsequent sources are selected conditional on the set of inflection classes $\mathbb{P}$ could belong to given the sources realized so far. Hence, the number of sources required per paradigm is upper bounded by the static approach and lower bounded by the dynamic.

Our FPC approach is a neural update, inspired by their adaptive approach. While their implementation tracks viable inflection classes explicitly with rules operating on oracularly segmented affixes, we use sequence models operating on whole words to remove reliance on oracular segmentation and leverage stem-internal phonology known to correlate with inflection class (Aronoff, 1992; Dressler and Thornton, 1996; Dawdy-Hesterberg and Pierrehumbert, 2014).

3 Frugal Paradigm Completion

This section describes the interactions of the three FPC components. As illustrated in Figure 1, the predictor takes a source cell and its realizing form as input, e.g., 3.SG.PRES: *sings*, or *cell 2*: *form 2* in the figure. The predictor is composed of as many sub-predictors as there are cells in the paradigm, each of which is trained to predict the entire paradigm from one source cell's realization. Cell 2 in the paradigm is grayed out in the figure, as this was provided as input so it does not have to be predicted. The predicted paradigm is evaluated by the oracle. If there are no errors, we are done. Otherwise, based on previous sources, SSA chooses a new cell to be realized by the oracle and gives it to the predictor as the next source. Because cell 3 is chosen in the figure, sub-predictor 3 will be used to predict the paradigm going forward, and cells 2 and 3 will both be grayed out. The process continues like this until all cells have been correctly predicted by at least one sub-predictor.

Crucially, during inference, each test paradigm is empty, i.e., no realization has been seen during training and no source is available to inflect from
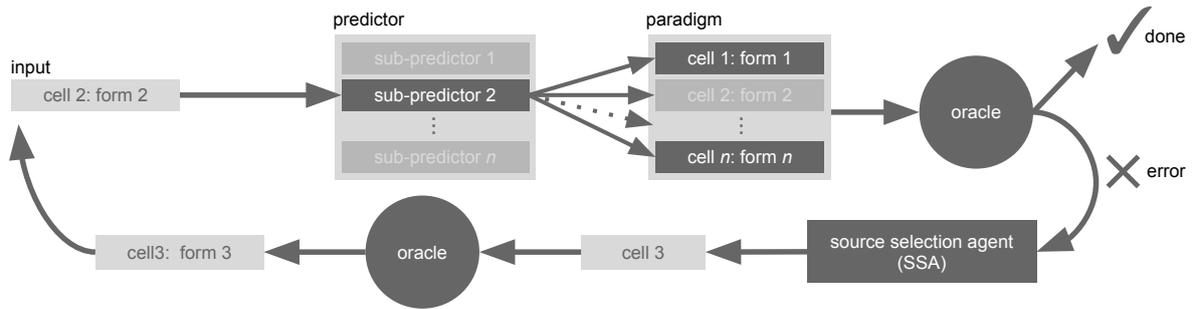
Figure 1: Schematic representation of the flow of Frugal Paradigm Completion at inference time.

a-priori. Our setup aims to minimize the number of sources which the SSA must request from the oracle (typically a human in the loop at inference time) to predict the remaining paradigm slots correctly.

3.1 Predictor

The predictor outputs a target form given its cell, a source form and the source form's cell as input. To train the predictor, for each possible source cell, we train a sub-predictor to predict every possible target form in every paradigm in the training data given the realization of that source cell in that paradigm. Details of all sequence model architectures are provided in Section 4.

3.2 Source Selection Agent

SSA's choice of a cell for a given paradigm depends on all previously selected cells for that paradigm and their corresponding forms. This allows SSA to learn, e.g., that given a previous English PST source, PST.PTCP should only be requested as a subsequent source if the PST form did not take the regular *-ed* suffix. Otherwise, PST.PTCP is likely to be regular and unlikely to contribute new information.

To induce such knowledge, we train SSA on an oracle policy of ideal source selections extracted from the train set (Ross et al., 2011; Ross and Bagnell, 2014; Welleck et al., 2019).[1] To extract the oracle policy, we divide the training lexicon into two folds and train one predictor on each, allowing us to cross-validate each predictor on its held out fold. For each training paradigm, we test which target forms can be correctly predicted by which source cells' sub-predictors. As shown for $\mathbb{SING}$

in Figure 2, we use this information to extract minimum set covers, i.e., the fewest source cells such that the union of the subsets they predict correctly covers the entire paradigm. These covers constitute the oracle policy used to train SSA.

The minimum set cover problem is NP-complete (Lund and Yannakakis, 1994; Kuhn et al., 2005), but we approximate it in $O(log_e|\mathbb{P}|)$ by iteratively selecting the cell whose subset most enlarges the union. We break ties by averaging predictiveness (Equation 1) over both folds, where fold F contains $|F|$ paradigms; $\mathbb{P}_m$, $|\mathbb{P}_m|$ cells; and $\text{Acc}(\mathbb{P}_m, \text{C}_{\text{trg}}, \text{C}_{\text{src}})$ returns 1 if using C_{src}'s realization as a source correctly predicts the form realizing cell C_{trg} in paradigm $\mathbb{P}_m$.

$$predictiveness(\text{C}_{\text{src}}, F) =$$
$$\frac{\sum_{m=1}^{|F|} \sum_{j=1}^{|\mathbb{P}_m|} \text{Acc}(\mathbb{P}_m, \text{C}_{\text{j}}, \text{C}_{\text{src}})}{\sum_{m=1}^{|F|} |\mathbb{P}_m|} \quad (1)$$

At this stage, paradigm covers are dynamic in that no single cell need be shared by all covers. Yet, when selecting the first source, SSA has no previous sources to condition on, making it impossible to predict the first cell. Thus, we get adaptive minimum set covers by designating the *start cell* to be that which occurs in the most dynamic covers. Then we re-approximate all covers such that each includes this cell.[2] Finally, we rank cells within each cover by the total number of covers in which they appear. For each cell in each cover, we train SSA to predict said cell from all higher ranked cells and their realizing forms (holding out 2% of them for development).

---

[1] While we borrow the term *oracle policy* from Imitation Learning (Ross et al., 2011; Ross and Bagnell, 2014; Welleck et al., 2019), we mimic the oracle policy with simple sequence learning. Our analysis suggests even this may be more machinery than necessary.

[2] We train and test on a single part-of-speech for each language, so each paradigm *should* contain the start cell. For defective paradigms lacking said cell, we back off to the most frequent cell that exists in the paradigm.

8250

Given	NFIN=*sing*	3.SG.PRES=*sings*	PRES.PTCP=*singing*	PST=*sang*	PST.PTCP=*sung*
Targets predicted correctly	NFIN=*sing* 3.SG.PRES=*sings* PRES.PTCP=*singing*	NFIN=*sing* 3.SG.PRES=*sings* PRES.PTCP=*singing*	NFIN=*sing* 3.SG.PRES=*sings* PRES.PTCP=*singing*	PST=*sang* PST.PTCP=*sung*	PST.PTCP=*sung*

Figure 2: Minimum set cover example for $\mathbb{SING}$, which is {NFIN, PST}.

3.3 Oracle

The oracle represents a human-in-the-loop during inference, providing requested source realizations to the predictor and informing SSA when a paradigm is complete and accurate (Figure 1). In our implementation, the oracle does not specify which individual predictions are incorrect, but it thus must resolve any discrepancies when two sub-predictors disagree after the fact. We do not attempt to model the additional cost this incurs, as it is unclear how to combine it with the presumably more expensive cost of correcting errors, which we model instead. This is worth re-visiting in future work.

4 Experimental Details

We evaluate 4 paradigm completion approaches on 7 languages. Here we discuss implementation, data and evaluation details.

4.1 Prediction Architecture

All sequence models in all implementations of any paradigm completion approach use the Transformer architecture (Vaswani et al., 2017). Here we describe the formatting of input and outputs as well as our hyperparameters.

Input and Output Formats Following Kann and Schütze (2016), input sequences combine characters and morphosyntactic features. The following is a sample input and output for a single source FPC sub-predictor specializing in the cell NFIN:

Input: *f l y* out_V.PTCP out_PST
Output: *f l o w n*

For any inflected-form-predicting sequence model whose input is not limited to realizations of a single cell—as in, e.g., the static principal parts approach—source cell features are prepended to the input as such:

Input: in_NFIN *f l y* out_V.PTCP out_PST
Output: *f l o w n*

For multi-source sequence models, the features of each source are inserted into the input and the target features are listed after the first source. We experimented with several different multi-source representations and the Transformer performed fairly similarly with all of them.

Input: in_NFIN *f l y* out_V.PTCP out_PST
        in_PST *f l e w*
Output: *f l o w n*

The FPC's SSA predicts not a form, but a cell, conditional on any previously realized sources. To predict the first source, it is given nothing and will thus deterministically select the best starting cell as determined by the oracle policy (see Section 3.2). To predict any subsequent source, it conditions on the realizations of all previously requested sources for that paradigm. The following exemplifies SSA inputs and outputs when predicting the second source for paradigm $\mathbb{FLY}$:

Input: in_NFIN *f l y*
Output: in_V.PTCP in_PST

Wu et al. (2018) and others have achieved improvements by embedding morphosyntactic features separately and concatenating them to the encoder output prior to feeding it to the decoder. Our error analysis, however, suggests Transformers handle Kann and Schütze (2016)-style input well. More sophisticated feature handling may not be necessary, but should be investigated in future work.

Hyperparameters We train all Transformer models for 100 epochs in batches of 64 with 0.1 dropout probability. The final model is restored from the epoch with the highest dev accuracy. We stop early if there is no improvement for 20

		Train	Dev	Test
Arabic	paradigms	1006	100	100
nouns	instances	24160	2260	2352
German	paradigms	1031	100	100
verbs	instances	27762	2690	2692
English	paradigms	2908	200	201
verbs	instances	14522	1000	1001
Russian	paradigms	3289	100	100
nouns	instances	37423	1133	1137
Latin	paradigms	1630	100	100
nouns	instances	19150	1180	1174
Hungarian	paradigms	1405	100	100
nouns	instances	47689	3400	3383
Irish	paradigms	549	100	100
nouns	instances	6460	1197	1195

Table 1: Number of paradigms and instances by split for every language and POS considered.

epochs. The only exception is during FPC cross-validation where sub-predictor models are trained for only 50 epochs with early stopping after 5 epochs without improvement. This is just to reduce computational cost as it is sufficient to induce an oracle policy. The final sub-predictor models however (those used at inference time, not those used to induce the oracle policy), are trained on the full training data set using the full 100 epochs with 20 epochs patience for early stopping. As for Transformer-specific hyperparameters, using the original notation of Vaswani et al. (2017), we set $N = 4$, $d_{model} = 128$, $d_{ff} = 512$, and $h = 8$, scaling down the hyperparameters recommended for machine translation as our task is less expensive (Aharoni and Goldberg, 2017; Wu et al., 2018).

4.2 Data Preparation

For every language and part of speech (POS) considered, we extract train, dev and test sets from UniMorph (Kirov et al., 2018). Each split contains full paradigms, though the cells realized in each may vary due to defectiveness (Corbett, 2005; Sims, 2015). We filter many gold errors by removing paradigms for which no realization can be attested in actual text. We use Universal Dependencies (UD) (Nivre et al., 2016) to check for attestations. We also filter overabundant realizations (multiple forms realizing one cell), keeping

only the most frequent form, as attested in UD. While some languages allow for overabundance (Thornton, 2010, 2011), in UniMorph, this often indicates a gold error.

We randomly divide paradigms into splits such that train is maximally large and dev and test contain at least 100 paradigms and 1,000 realizations. Exact quantities are displayed in Table 1. Arabic, German, English, and Russian were used for development, while Irish, Hungarian, and Latin were only evaluated after fixing hyperparameters. The languages considered represent 3 families and 4 diverse Indo-European branches. They exhibit multiple *non-canonical* behaviors (Corbett, 2005) and present diverse challenges from non-concatenative morphology to complex inflection class systems.

4.3 Evaluation

Paradigm completion is usually evaluated via exact match *accuracy* on held out target forms (Cotterell et al., 2016, 2017a, 2018; McCarthy et al., 2019). Yet we use as many sources as are necessary to reach 100% accuracy in predicting the remaining slots, so accuracy is not a meaningful metric for the FPC. Some theoretical works focus on the sources required to unambiguously complete a paradigm given some implicit knowledge of viable inflection classes (Finkel and Stump, 2007; Ackerman and Malouf, 2013). Yet these tend not to propose actual paradigm completion models or evaluate their decisions in ambiguous cases. To evaluate our system and bridge these traditions, we propose *auto-rate*:

$$auto\text{-}rate = \frac{\sum_{i=1}^{n} auto(\mathbb{P}_i)}{\sum_{i=1}^{n} |\mathbb{P}_i|}, \qquad (2)$$

where $auto(\mathbb{P})$ denotes the number of realizations correctly predicted while not having been provided as sources for paradigm $\mathbb{P}$ by the oracle.

Intuitively, auto-rate is like accuracy but it counts oracularly provided sources as additional errors since both errors and sources require labor, i.e., sources require manual input and errors, post-correction. We also report *manual cells per paradigm*, i.e., sources plus errors. Of course, FPC resolves all errors eventually, but other systems can make errors requiring post-correction.

4.4 Baselines

We compare the FPC method to three baselines. One is a variant of FPC using a *random* SSA.

This allows us to distinguish the benefit of a smart SSA from that of simply receiving additional feedback from an oracle in the loop. Each time a source must be selected, random SSA chooses randomly without replacement. Its performance is averaged over two runs. The *lemma* approach baseline predicts all paradigm forms from one designated source: the lemma. Finally, for the *static* approach baseline, we considered two *static* approach implementations. The single-source implementation predicts each target from the source that is, in theory, its best predictor (Kann and Schütze, 2018). The multi-source implementation concatenates these sources, predicting each target from the concatenated input. As results are nearly identical for either implementation, we report results only for single-source—with the exception of Latin, as explained presently.

For some languages, there is little theoretical or pedagogical literature to help identify the best sources for the static approach. Our single-source static approach for Arabic nouns predicts singular and dual forms from SG;NDEF;NOM and plurals from PL;NDEF;NOM. In theory, any non-plural plus any plural should be sufficient (Brustad et al., 2005; Habash, 2010). For German verbs, we predict present and imperative forms from NFIN and past forms from IND;PST;1SG (Grebe et al., 1966). We predict English present forms from NFIN; PST and V.PTCP;PST predict themselves. For Russian nouns, Zaliznyak (1980) argues for five sources, yet Parker (2016) demonstrates that three are usually sufficient. We follow the latter, predicting all nominative or accusative forms from ACC;SG, all other singulars from INS;SG, and all other plurals from GEN;PL. In preliminary experiments, we found this to match the accuracy of the five source approach, thus achieving a higher auto-rate. For Latin, we could not evaluate a single-source static implementation as it is unclear which source cell best predicts each target. The multi-source static approach for Latin nouns predicts all forms from NOM;SG and GEN;SG (following the classical grammatical analyses of Varro, Priscian and the Roman *ars grammatica*). For Irish and Hungarian, we do not evaluate a static approach as we lack the requisite linguistic knowledge to determine the best sources.

	Accuracy		Auto-rate		Mcpp	
	Dev	Test	Dev	Test	Dev	Test
Arabic nouns						
Lemma	62.0	58.8	59.3	56.5	9.6	10.7
Static	95.9	99.4	89.5	93.1	2.9	2.1
Random Ag.			90.2	90.9	**2.2**	2.2
FPC			90.2	93.6	2.2	**1.5**
German verbs						
Lemma	87.6	89.0	84.1	85.8	4.3	4.0
Static	94.1	96.4	86.7	88.9	3.6	3.0
Random Ag.			90.0	92.1	2.4	1.9
FPC			**91.8**	92.5	2.0	1.8
English verbs						
Lemma	96.5	94.0	76.7	74.2	1.2	**1.3**
Static	99.7	98.4	39.7	38.4	3.0	3.0
Random Ag.			76.0	73.3	1.2	1.4
FPC			**77.3**	**74.3**	1.1	1.3
Russian nouns						
Lemma	97.1	95.6	88.3	87.5	**1.3**	1.5
Static	98.4	98.3	72.6	72.3	3.2	3.2
Random Ag.			86.1	84.3	1.6	1.8
FPC			**88.5**	**89.1**	1.3	**1.2**
Latin nouns						
Lemma	65.5	51.6	63.6	49.6	5.1	6.7
Static	97.7	96.8	80.8	79.7	2.3	2.4
Random Ag.			85.9	84.7	1.7	1.8
FPC			**89.0**	**87.8**	1.3	1.4
Hungarian nouns						
Lemma	95.6	90.9	92.8	88.0	2.5	4.1
Random Ag.			95.0	94.6	1.7	1.9
FPC			**95.5**	**95.2**	1.5	1.6
Irish nouns						
Lemma	63.5	66.9	56.1	59.6	5.4	5.0
Random Ag.			64.9	68.2	4.2	3.8
FPC			**72.1**	**69.6**	3.3	3.6

Table 2: Evaluation of paradigm completion approaches with metrics defined in Section 4. We do not report accuracy for FPC or its random agent variant (Random Ag.), as it is trivially 100% (see Section 4.3). Mcpp stands for *Manual cells per paradigm*.

5 Results and Discussion

As shown in Table 2, FPC always ties or beats the next best approach, while the next best approach varies by language. On average, FPC reduces labor by 63% over the lemma approach, 47% over static, 16% over random agent, and 13% over the next best approach. Its success is mainly due to (1) making predictions from fewer sources than are required for fully disambiguating inflection class and (2) receiving feedback after each source.

Surprisingly, training a sophisticated SSA does not improve much over using a random agent. We

argue this is due to an unexpectedly large margin of error in the agent's source selection task. Despite the complexity of source selection strategies required for inflection class disambiguation, FPC uses lexical frequencies to expect regularity and stem-internal clues to anticipate irregular classes, requiring a median of just one source per paradigm for all languages except under-resourced Irish. Furthermore, inspection of the source selection minimum set covers reveals that it is often the case that a paradigm can be completed correctly from any single source. This is surprising in light of the precise strategies required for completely deterministic paradigm completion in Finkel and Stump (2007)'s framework and in light of Albright (2002)'s case for the privileged status of a single form per paradigm, though in our framework with full words and full paradigms for training, it seems that many sources can often serve as *good enough* singleton principal parts. This supports Bonami and Beniamine (2016) proposal of gradient principal part analyses.

6 Analysis

Here, we discuss patterns relating SSA's first and second sources chosen (Figures 3a-b and 4a-b) to the inter-predictability of cells represented by heat maps (3c and 4c). Maps display the average accuracies with which each target (column) can be predicted from each source (row). We analyze specific SSA choices and predictor errors in Arabic and Latin.

The maps (for all languages, see the Appendix) suggest complexity can be distributed within paradigms in systematically distinct ways. Ackerman and Malouf (2013) propose integrative (I-) complexity, using average conditional entropy to describe paradigmatic organization, but this has been criticized for obscuring differences in the predictability of sub-paradigm regions (Cotterell et al., 2019; Elsner et al., 2019). To remedy this, we propose a typology for measuring the extent to which I-complexity is realized via different organizational strategies, which is useful for discussing source selection strategies. Our typology describes paradigms in terms of *mutual predictability*, the correlation of a map and its transpose, and *entropy predictiveness*, the negative correlation of cells' average predictiveness (see Equation 1) and average predictability, defined here in comparable terms as:

$$predictability(c_{\text{trg}}, F) =$$

$$\frac{\sum_{m=1}^{|F|} \sum_{j=1}^{|\mathbb{P}_m|} \text{Acc}(\mathbb{P}_m, c_{\text{trg}}, c_{\text{j}})}{\sum_{m=1}^{|F|} |\mathbb{P}_m|} \quad (3)$$

Intuitively, a paradigm is mutually predictable if the fact that cell A predicts cell B means that B is likely to predict A. Such paradigms often feature regions of mutually predictable cells (as in 3c), such that an optimal strategy avoids picking multiple sources from one region. For entropy predictive paradigms, if A is generally more difficult to predict than B, A is likely to be a better predictor of the remaining cells (following the information theoretic logic that surprisal is informative (Shannon, 1948; Jaeger, 2010)). For such paradigms, the optimal strategy selects the source which would have been the most difficult target to predict.

Unlike Sims (2020)'s graph-theoretic typology for describing inflection class structure, our typology is a two-dimensional description of how the optimal paradigm completion strategy is affected by underlying class structure. In this sense, our typology is complementary to hers and future work might investigate the relationship between traits in her typology and mutual predictability or entropy predictiveness. Furthermore, our typology might be updated to consider the impact of type frequency (Sims and Parker, 2016) in a framework where distributional data is available.

Figure 5 demonstrates that cross-linguistic variation is vast with respect to our typology, as some languages even exhibit negative entropy predictiveness or mutual predictability. This partly explains why non-FPC approaches perform erratically: if paradigmatic organization varies by language, source selection strategies must be able to adapt to the data.

6.1 Arabic Error Analysis

Arabic nouns are mutually predictable (Figure 5). Any singular or dual form can predict another. Plural forms also predict each other. Yet, in general, plurals are less predictive/able (Figure 3c) due to several inflection classes varying in the plural. The *sound* plurals take suffixes while *broken* plural classes are realized via non-concatenative processes. For example, راكب *rAkb*, *rider* from root ر ك ب *r k b*, takes the broken plural pattern _ _ A _, becoming ركاب *rkAb*. Yet, having heard only singular realizations, a human might posit a sound plural, i.e., *راكبون* *rAkbwn*, realizing the
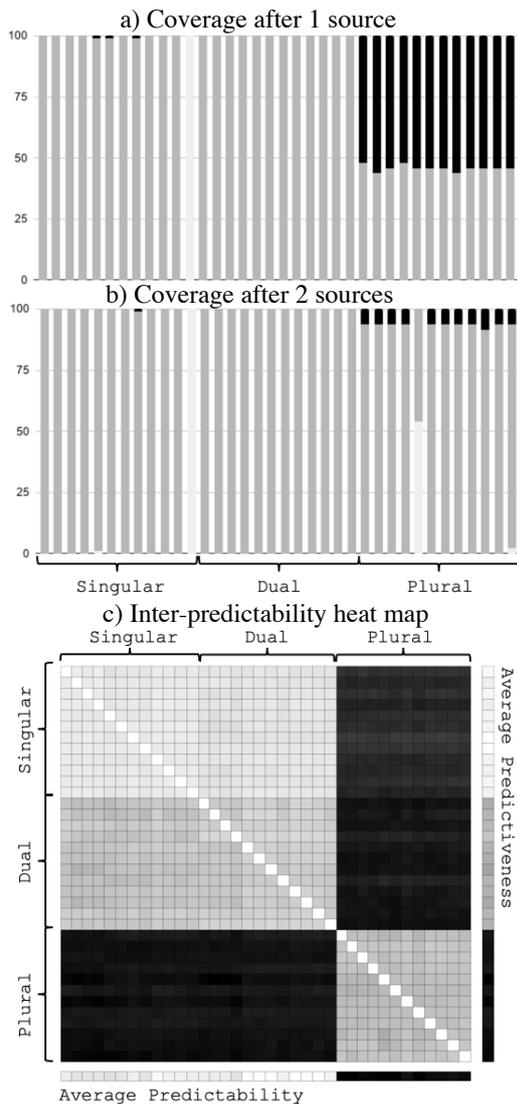
Figure 3: Arabic analysis. (a) and (b) define how likely each cell is to be a source (white), correctly predicted (gray), or an error (black) after one (a) or two (b) sources. (c) shows the predictiveness/ability of source (rows) and target (columns) cells. Darker cells are less predictive/able. For a more detailed rendering of this graphic, please see the appendix.

Figure 4: Latin analysis. (a) and (b) define how likely each cell is to be a source (white), correctly predicted (gray), or an error (black) after one (a) or two (b) sources. (c) shows the predictiveness/ability of source (rows) and target (columns) cells. Darker cells are less predictive/able. For a more detailed rendering of this graphic, please see the appendix.

more productive exponent.

SSA learns an ideal strategy, requesting a singular source (Figure 3a) and then a plural (3b). Interestingly, 6 of 18 sound feminine plurals (most frequent single class) require multiple sources and 8 of 28 broken plurals do not. Thus, the predictor does not default to regularity, but uses stem-internal phonology to anticipate irregularity. Most errors made from the first source posit a viable broken plural, just not the right one. In future work, modeling semantics can fix such errors, e.g., knowing that راكب *rAkb* is animate makes plural

*رواكب *rwAkb* unlikely, as animate nouns seldom take that inflection class.

For future work, we can pre-train on raw corpora to give our model access to such information (Devlin et al., 2019). Indeed Erdmann and Habash (2018) found distributional information to benefit inflectional paradigm clustering in Arabic. Though the benefits should generalize as semantics correlates with inflection class in many languages (Wurzel, 1989; Aronoff, 1992; Harris, 1992; Noyer, 1992; Carstairs-McCarthy, 1994; Corbett and Fraser, 2000; Kastner, 2019).
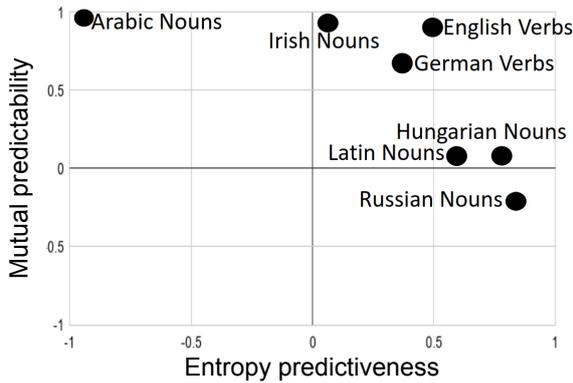
Figure 5: Integrative typology for describing paradigms by the mutual predictability of cells and the predictive power of hard-to-predict cells.

Cell	1st	2^{nd}M	2^{nd}N	3rd	4^{th}M	4^{th}N
NOM;SG	a	us	um	*varies*	us	ū
GEN;SG	ae	ī	ī	is	ūs	ūs
DAT;SG	ae	ō	ō	ī	uī	ū
ACC;SG	am	um	um	em	um	ū
ABL;SG	ā	ō	ō	em	ū	ū
VOC;SG			*generally matches NOM;SG*			
NOM;PL	ae	ī	a	ēs	ūs	ua
GEN;PL	arum	ōrum	ōrum	um	uum	uum
DAT;PL	is	īs	īs	ibus	ibus	ibus
ACC;PL	as	ōs	a	ēs	ūs	ua
ABL;PL	is	īs	īs	ibus	ibus	ibus
VOC;PL			*generally matches NOM;PL*			

Table 3: Plat of the suffixes taken for each Latin nominal declension.

6.2 Latin Error Analysis

Latin is not mutually predictable with moderate entropy predictiveness. SSA's choices are, at first, opaque, but Table 3 shows that ACC;PL narrows the inflection class to variants of one declension. Remaining ambiguity mostly involves 3rd declension nominative and vocative realizations, which can usually be predicted from the preferred second source cell, VOC;SG. 44 of 100 test paradigms were 3rd declension, which required multiple sources at the highest rate (16 of 44; 2nd masculine declension was next highest at 3 of 15). There was no correlation between declension and second source chosen, yet high auto-rate suggests SSA's choices may not need to condition on previously realized source forms, but only their cells.

While 77 of 100 paradigms were completed

from a single source, we found paradigms requiring three sources that might be completable from two using a multi-source FPC implementation. For example, *gregēs*, *flocks* realizes $\mathbb{GREX}$.ACC;PL, but the predictor mistakenly posits *gregium for GEN;PL from this source, guessing the wrong 3rd declension variant. While second source VOC;SG *grex* corrects this, it obscures the underlying stem, as *x* can be an allophone of *g* or *c*. Thus, we still get an error, *grecum. A multi-source predictor could avoid forgetting the underlying allophone *g* after seeing the second source.[3] That said, multi-source FPC is not as simple as multi-source static. Heuristic sampling of training instances based on the oracle policy yields predictors that only attend to one source or make bad predictions when only given one. This is worth exploring further in future work as there is more evidence of paradigms that are difficult to handle without jointly encoding sources in the linguistic literature (Corbett, 2005; Bonami and Beniamine, 2016).

7 Conclusion

We presented Frugal Paradigm Completion, which reduces the manual labor required to expand a morphological lexicon by 16-63% over competitive approaches across 7 languages. We demonstrated that typologically distinct morphological systems require unique treatment and benefit from our SSA, that learns its strategy from data. We found that inducing this strategy is not as challenging as previously suggested (Finkel and Stump, 2007). Thus, SSA might be replaced with a less costly architecture while our model might be improved by conditioning on semantics and jointly decoding from a variable number of sources.

---

[3]See *rex–regis*, *king* or *pax–pacis*, *peace*, which are technically conditioned on preceding vowel quality, though there are probably not enough training examples for the model to learn that.

References

Farrell Ackerman, James P Blevins, and Robert Malouf. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. *Analogy in grammar: Form and acquisition*, pages 54–82.

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, pages 429–464.

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015.

Thomas Ahern. 1990. The role of linguistics in Latin instruction. *The Classical World*, pages 505–510.

Adam Albright. 2002. *The identification of bases in morphological paradigms. UCLA Ph.D.* Ph.D. thesis, dissertation.

Mark Aronoff. 1992. Noun classes in Arapesh. In *Yearbook of Morphology 1991*, pages 21–32. Springer.

Sacha Beniamine, Olivier Bonami, and Benoît Sagot. 2018. Inferring inflection classes with description length. *Journal of Language Modelling*, 5(3):465–525.

Olivier Bonami and Sarah Beniamine. 2016. Joint predictiveness in inflectional paradigms. *Word Structure*, 9(2):156–182.

Kristen Brustad, Abbas Al-Tonsi, and Mahmoud Al-Batal. 2005. *Al-Kitaab Fii Ta'Aallum Al-'Arabiyya: A Textbook for Arabic*. Georgetown University Press.

Tim Buckwalter. 2002. Arabic morphological analyzer (AraMorph). *Linguistic Data Consortium, Philadelphia*.

Andrew Carstairs-McCarthy. 1994. Inflection classes, gender, and the principle of contrast. *Language*, pages 737–788.

Greville G Corbett. 2005. The canonical approach in typology. *Linguistic diversity and language theories*, 25:49.

Greville G Corbett and Norman M Fraser. 2000. Gender assignment: a typology and a model. *Systems of nominal classification*, 4:293–325.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 2018. The conll–sigmorphon 2018 shared task: Universal morphological reinflection. *Proceedings of the CoNLL SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017a. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. *CoRR*, abs/1706.09031.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the Workshop of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON)*, Berlin, Germany.

Ryan Cotterell, John Sylak-Glassman, and Christo Kirov. 2017b. Neural graphical models over strings for principal parts morphological paradigm completion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 759–765.

Lisa Garnand Dawdy-Hesterberg and Janet Breckenridge Pierrehumbert. 2014. Learnability and generalisation of Arabic broken plural nouns. *Language, cognition and neuroscience*, 29(10):1268–1282.

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.

Wolfgang U Dressler and Anna M Thornton. 1996. Italian nominal inflection. *Wiener Linguistische Gazette*, pages 55–57.

Markus Dreyer and Jason Eisner. 2011. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 616–627. Association for Computational Linguistics.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1185–1195.

Micha Elsner, Andrea D Sims, Alexander Erdmann, Antonio Hernandez, Evan Jaffe, Lifeng Jin, Martha Booker Johnson, Shuan Karim, David L King, Luana Lamberti Nunes, et al. 2019. Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute? *Journal of Language Modelling*, 7(1):53–98.

Alexander Erdmann and Nizar Habash. 2018. Complementary strategies for low resourced morphological modeling. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 54–65.

Ramy Eskander, Nizar Habash, and Owen Rambow. 2013. Automatic extraction of morphological lexicons from morphologically annotated corpora. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1032–1043.

Raphael Finkel and Gregory Stump. 2007. Principal parts and morphological typology. *Morphology*, 17(1):39–75.

Markus Forsberg, Harald Hammarström, and Aarne Ranta. 2006. Morphological lexicon extraction from raw text data. In *International Conference on Natural Language Processing (in Finland)*, pages 488–499. Springer.

Paul Grebe, Helmut Gipper, and Günther Drosdowski, editors. 1966. *Duden Grammatik der deutschen Gegenwartssprache*, 2., verm. und verb. aufl edition. Bibliogr. Inst., Dudenverl., Mannheim.

Wren Jones Grinstead. 1916. *An Introduction to the Psychology and Pedagogy of High School Latin: With Special Reference to the Value of a One-year Course*, volume 1. University of Wisconsin–Madison.

Nizar Y Habash. 2010. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

James W Harris. 1992. The form classes of Spanish substantives. In *Yearbook of morphology 1991*, pages 65–88. Springer.

T Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62.

Katharina Kann and Hinrich Schütze. 2016. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 62–70.

Katharina Kann and Hinrich Schütze. 2018. Neural transductive learning and beyond: Morphological generation in the minimal-resource setting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3254–3264.

Itamar Kastner. 2019. Templatic morphology as an emergent property. *Natural Language & Linguistic Theory*, 37(2):571–619.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian J Mielke, Arya McCarthy, Sandra Kübler, et al. 2018. UniMorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Fabian Kuhn, Pascal Von Rickenbach, Roger Wattenhofer, Emo Welzl, and Aaron Zollinger. 2005. Interference in cellular networks: The minimum membership set cover problem. In *International Computing and Combinatorics Conference*, pages 188–198. Springer.

Carsten Lund and Mihalis Yannakakis. 1994. On the hardness of approximating minimization problems. *Journal of the ACM (JACM)*, 41(5):960–981.

Arya D McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sebastian Mielke, Jeffrey Heinz, et al. 2019. The SIGMORPHON 2019 shared task: Crosslinguality and context in morphology. In *Proceedings of the 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Fabio Montermini and Olivier Bonami. 2013. Stem spaces and predictability in verbal inflection. *Lingue e linguaggio*, 12(2):171–190.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.

Robert Rolf Noyer. 1992. *Features, positions and affixes in autonomous morphological structure*. Ph.D. thesis, Massachusetts Institute of Technology.

Jeff Parker. 2016. *Inflectional Complexity and Cognitive Processing: An Experimental and Corpus-based Investigation of Russian Nouns*. Ph.D. thesis, Ohio State University.

Stephane Ross and J Andrew Bagnell. 2014. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635.

Benoît Sagot. 2010. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*.

Wolfgang Seeker and Jonas Kuhn. 2013. Morphological and syntactic case in statistical dependency parsing. *Computational Linguistics*, 39(1):23–55.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Andrea D Sims. 2015. *Inflectional defectiveness*, volume 148. Cambridge University Press.

Andrea D Sims. 2020. Inflectional networks: Graph-theoretic tools for inflectional typology. *Proceedings of the Society for Computation in Linguistics: Vol*, 3:10.

Andrea D Sims and Jeff Parker. 2016. How inflection class systems work: On the informativity of implicative structure. *Word Structure*, 9(2):215–239.

Gregory Stump and Raphael A Finkel. 2013. *Morphological typology: From word to paradigm*, volume 138. Cambridge University Press.

Marko Tadić and Sanja Fulgosi. 2003. Building the Croatian morphological lexicon. In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, pages 41–46. Association for Computational Linguistics.

Anna Thornton. 2011. Overabundance (multiple forms realizing the same cell): a non-canonical phenomenon in Italian verb morphology.

Anna M Thornton. 2010. Towards a typology of overabundance. In *Décembrettes 7: International Conference on Morphology, University of Toulouse*, pages 2–3.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vincent Wan, Chun an Chan, Tom Kenter, Jakub Vit, and Rob Clark. 2019. CHiVE: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network. In *ICML 2019*.

Sean Welleck, Kianté Brantley, Hal Daumé III, and Kyunghyun Cho. 2019. Non-monotonic sequential text generation. In *International Conference on Machine Learning*, pages 6716–6726.

Shijie Wu and Ryan Cotterell. 2019. Exact hard monotonic attention for character-level transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537.

Shijie Wu, Pamela Shapiro, and Ryan Cotterell. 2018. Hard non-monotonic attention for character-level transduction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4425–4438.

Wolfgang Ullrich Wurzel. 1989. *Inflectional morphology and naturalness*, volume 9. Springer Science & Business Media.

A. A. Zaliznyak. 1980. *The grammar dictionary of the Russian language*.

Heiga Zen, Yannis Agiomyrgiannakis, Niels Egberts, Fergus Henderson, and Przemysław Szczepaniak. 2016. Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*.

A Expanded Results

The figures in this appendix demonstrate the coverage after 1 and 2 sources for every language considered as well as their inter-predictability heat maps. Figures are enlarged to show all individual cells for the reader's convenience. Hence, the Arabic and Latin figures in this appendix correspond to Figures 3 and 4 in Section 6, but show more detail.
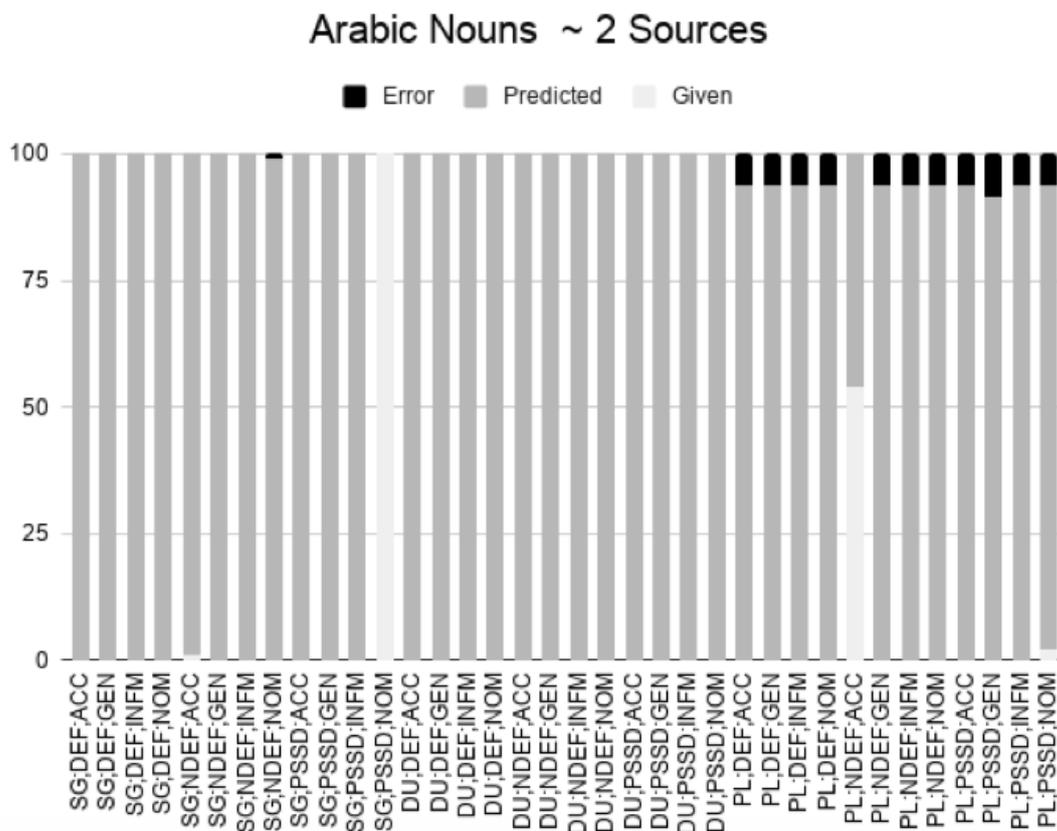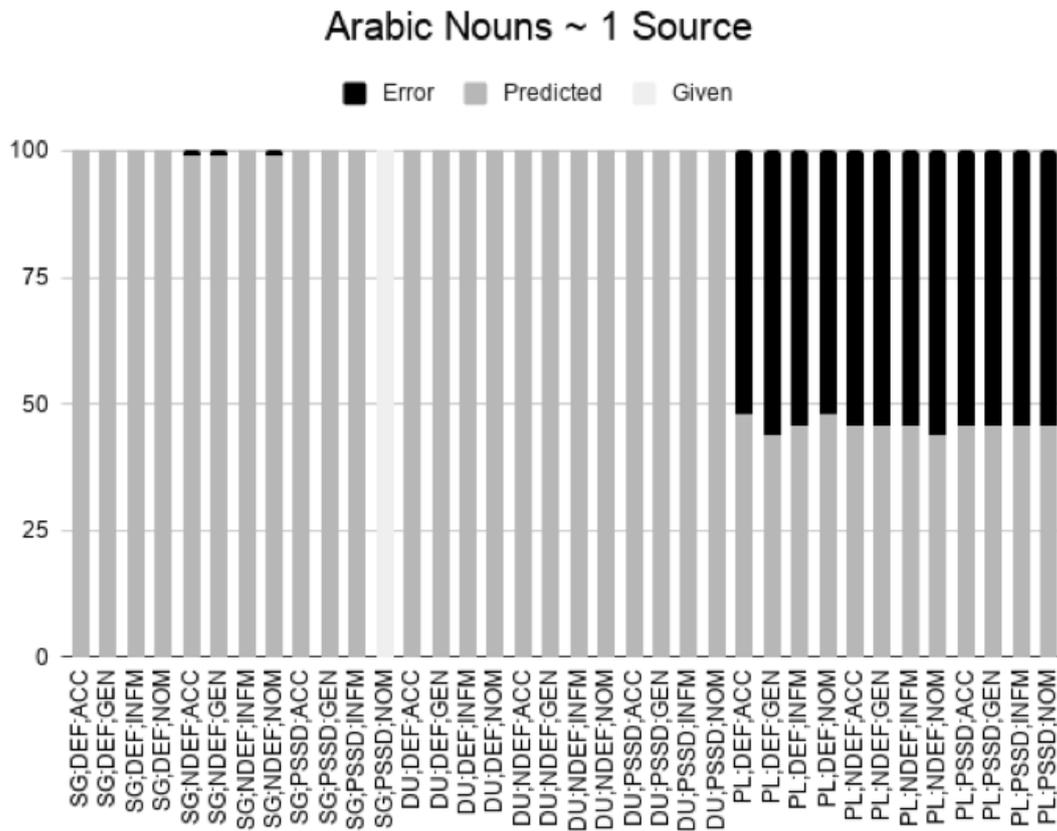
Figure 6: Coverage of Arabic target cells after SSA chooses the first two sources.

Arabic Noun Inter-Predictability

Figure 7: Inter-predictability heat map of Arabic cells.

Figure 8: Coverage of German target cells after SSA chooses the first two sources.

German Verb Inter-Predictability

|	NFIN	IMP;2;PL	IMP;2;SG	IND;PRS;1;SG	IND;PRS;2;SG	IND;PRS;3;SG	SBJV;PRS;1;SG	SBJV;PRS;2;SG	SBJV;PRS;3;SG	IND;PRS;1;PL	IND;PRS;2;PL	IND;PRS;3;PL	SBJV;PRS;1;PL	SBJV;PRS;2;PL	SBJV;PRS;3;PL	IND;PST;1;SG	IND;PST;2;SG	IND;PST;3;SG	SBJV;PST;1;SG	SBJV;PST;2;SG	SBJV;PST;3;SG	IND;PST;1;PL	IND;PST;2;PL	IND;PST;3;PL	SBJV;PST;1;PL	SBJV;PST;2;PL	SBJV;PST;3;PL	Average Predictiveness
NFIN	1.00	0.82	0.78	0.83	0.77	0.77	0.84	0.83	0.84	0.83	0.82	0.83	0.83	0.83	0.83			0.71			0.71			0.73			0.77	
IMP;2;PL	0.87	1.00	0.87	0.91	0.86	0.87	0.93	0.93	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.79	0.78	0.79	0.79	0.79	0.79	0.78	0.78	0.78	0.79	0.79	0.79	0.86
IMP;2;SG	0.81	0.87	1.00	0.87	0.87	0.87	0.88	0.87	0.88	0.87	0.87	0.87	0.87	0.87	0.87	0.75	0.75	0.75	0.76	0.76	0.76	0.75	0.75	0.75	0.76	0.76	0.76	0.82
IND;PRS;1;SG	0.86	0.89	0.85	1.00	0.79	0.82	0.91	0.91	0.91	0.91	0.89	0.91	0.91	0.91	0.91	0.75	0.75	0.75	0.76	0.76	0.76	0.76	0.75	0.76	0.76	0.76	0.76	0.83
IND;PRS;2;SG	0.80	0.85	0.84	0.83	1.00	0.90	0.85	0.86	0.85	0.84	0.85	0.84	0.84	0.85	0.84	0.77	0.77	0.77	0.76	0.77	0.77	0.78	0.77	0.77	0.77	0.76	0.77	0.82
IND;PRS;3;SG	0.82	0.88	0.87	0.86	0.90	1.00	0.87	0.87	0.87	0.86	0.87	0.86	0.86	0.88	0.86	0.77	0.76	0.77	0.77	0.77	0.77	0.78	0.77	0.78	0.78	0.77	0.78	0.83
SBJV;PRS;1;SG	0.84	0.90	0.82	0.88	0.82	0.83	1.00	0.90	0.91	0.89	0.90	0.89	0.89	0.91	0.89	0.74	0.75	0.75	0.77	0.77	0.77	0.75	0.75	0.75	0.77	0.77	0.77	0.83
SBJV;PRS;2;SG	0.87	0.90	0.85	0.89	0.84	0.83	0.91	1.00	0.91	0.90	0.89	0.90	0.90	0.91	0.90	0.77	0.76	0.77	0.78	0.77	0.78	0.77	0.76	0.77	0.78	0.77	0.78	0.84
SBJV;PRS;3;SG	0.84	0.88	0.83	0.87	0.81	0.81	0.89	0.89	1.00	0.89	0.88	0.89	0.89	0.90	0.89	0.74	0.73	0.74	0.75	0.74	0.75	0.74	0.73	0.74	0.75	0.74	0.75	0.82
IND;PRS;1;PL	0.86	0.89	0.85	0.89	0.82	0.83	0.91	0.91	0.91	1.00	0.89	0.91	0.91	0.91	0.91	0.76	0.75	0.76	0.76	0.76	0.77	0.75	0.75	0.76	0.76	0.77	0.77	0.83
IND;PRS;2;PL	0.86	0.92	0.85	0.90	0.83	0.84	0.92	0.91	0.91	0.90	1.00	0.90	0.91	0.92	0.90	0.79	0.78	0.79	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.85
IND;PRS;3;PL	0.88	0.90	0.86	0.90	0.82	0.83	0.92	0.92	0.92	0.92	0.90	1.00	0.92	0.92	0.92	0.77	0.76	0.77	0.78	0.78	0.78	0.78	0.77	0.77	0.78	0.79	0.78	0.85
SBJV;PRS;1;PL	0.88	0.90	0.85	0.90	0.82	0.84	0.92	0.92	0.92	0.92	0.90	0.92	1.00	0.92	0.92	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.84
SBJV;PRS;2;PL	0.86	0.90	0.83	0.88	0.83	0.84	0.91	0.91	0.91	0.90	0.90	0.89	0.90	1.00	0.90	0.74	0.74	0.74	0.74	0.74	0.75	0.74	0.74	0.74	0.74	0.74	0.75	0.82
SBJV;PRS;3;PL	0.85	0.89	0.84	0.88	0.81	0.82	0.90	0.90	0.90	0.90	0.89	0.90	0.90	0.89	1.00	0.73	0.72	0.72	0.74	0.74	0.74	0.73	0.73	0.74	0.74	0.74	0.74	0.82
IND;PST;1;SG	0.79	0.83	0.81	0.83	0.83	0.83	0.84	0.84	0.84	0.83	0.83	0.83	0.83	0.84	0.83	1.00	0.90	0.90	0.89	0.89	0.89	0.91	0.90	0.91	0.90	0.90	0.90	0.86
IND;PST;2;SG	0.80	0.84	0.80	0.81	0.73	0.83	0.84	0.84	0.85	0.84	0.83	0.84	0.84	0.84	0.84	0.89	1.00	0.90	0.90	0.90	0.90	0.91	0.91	0.91	0.90	0.90	0.90	0.86
IND;PST;3;SG	0.78	0.84	0.82	0.82	0.83	0.83	0.84	0.84	0.83	0.83	0.84	0.83	0.83	0.84	0.83	0.90	0.90	1.00	0.89	0.89	0.89	0.90	0.90	0.90	0.89	0.89	0.89	0.86
SBJV;PST;1;SG	0.77	0.84	0.83	0.82	0.81	0.82	0.84	0.83	0.84	0.83	0.84	0.83	0.83	0.84	0.83	0.88	0.88	0.89	1.00	0.91	0.92	0.90	0.90	0.90	0.92	0.92	0.92	0.86
SBJV;PST;2;SG	0.76	0.83	0.79	0.81	0.81	0.81	0.82	0.83	0.83	0.82	0.83	0.82	0.82	0.83	0.82	0.88	0.88	0.88	0.91	1.00	0.90	0.89	0.89	0.91	0.91	0.91		0.85
SBJV;PST;3;SG	0.80	0.84	0.81	0.83	0.81	0.83	0.85	0.84	0.85	0.84	0.84	0.84	0.84	0.85	0.84	0.89	0.90	0.89	0.92	0.92	1.00	0.90	0.90	0.90	0.92	0.92	0.92	0.87
IND;PST;1;PL	0.79	0.85	0.81	0.81	0.84	0.84	0.84	0.84	0.84	0.83	0.84	0.83	0.83	0.84	0.83	0.92	0.93	0.92	0.92	0.91	0.92	1.00	0.94	0.94	0.92	0.92	0.91	0.87
IND;PST;2;PL	0.76	0.83	0.79	0.79	0.81	0.83	0.82	0.82	0.82	0.81	0.82	0.81	0.81	0.82	0.81	0.88	0.90	0.88	0.88	0.88	0.88	0.90	1.00	0.90	0.88	0.88	0.88	0.85
IND;PST;3;PL	0.78	0.83	0.81	0.82	0.82	0.83	0.84	0.83	0.84	0.82	0.83	0.83	0.82	0.84	0.83	0.89	0.90	0.89	0.88	0.88	0.88	0.91	0.91	1.00	0.88	0.88	0.88	0.86
SBJV;PST;1;PL	0.81	0.85	0.81	0.82	0.82	0.84	0.85	0.84	0.85	0.83	0.84	0.83	0.83	0.85	0.83	0.89	0.89	0.90	0.92	0.91	0.92	0.90	0.91	0.91	1.00	0.92	0.92	0.87
SBJV;PST;2;PL	0.78	0.83	0.79	0.80	0.81	0.82	0.82	0.83	0.83	0.81	0.82	0.81	0.81	0.82	0.82	0.86	0.87	0.87	0.89	0.89	0.89	0.87	0.87	0.87	0.89	1.00	0.89	0.85
SBJV;PST;3;PL	0.81	0.86	0.83	0.83	0.83	0.85	0.85	0.86	0.85	0.84	0.86	0.84	0.84	0.86	0.84	0.88	0.89	0.88	0.92	0.92	0.92	0.90	0.90	0.90	0.91	0.92	1.00	0.87
Average Predictability	0.83	0.87	0.83	0.85	0.83	0.84	0.87	0.87	0.88	0.87	0.87	0.87	0.87	0.87	0.87				0.83	0.83	0.83	0.82	0.82	0.83	0.83	0.83	0.83	

Figure 9: Inter-predictability heat map of German cells.
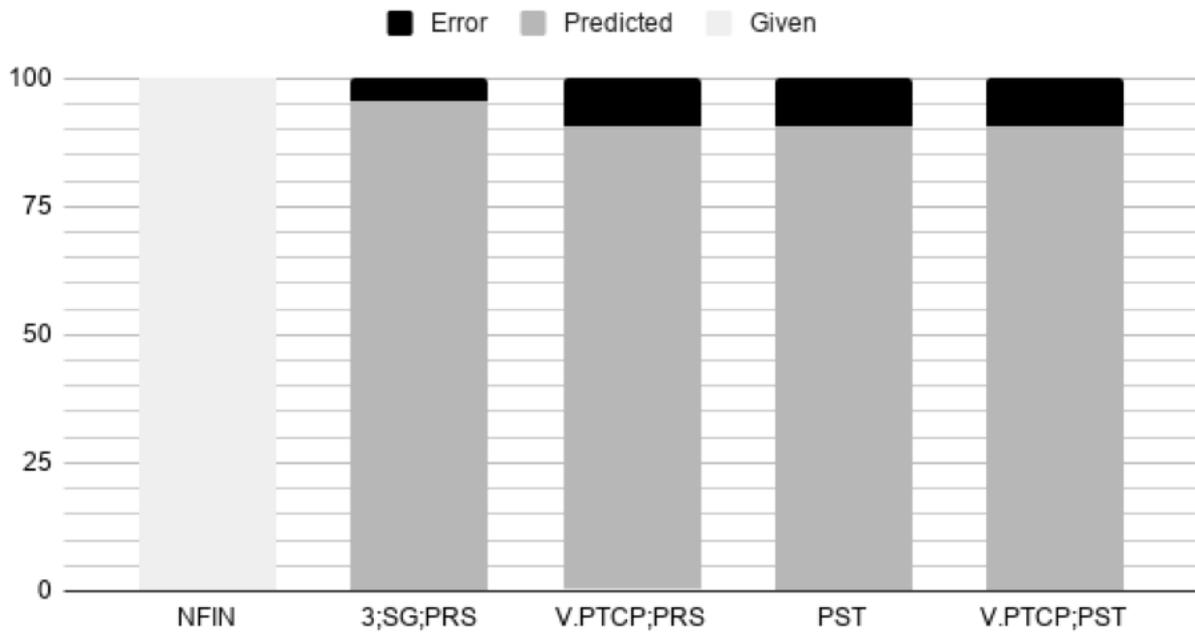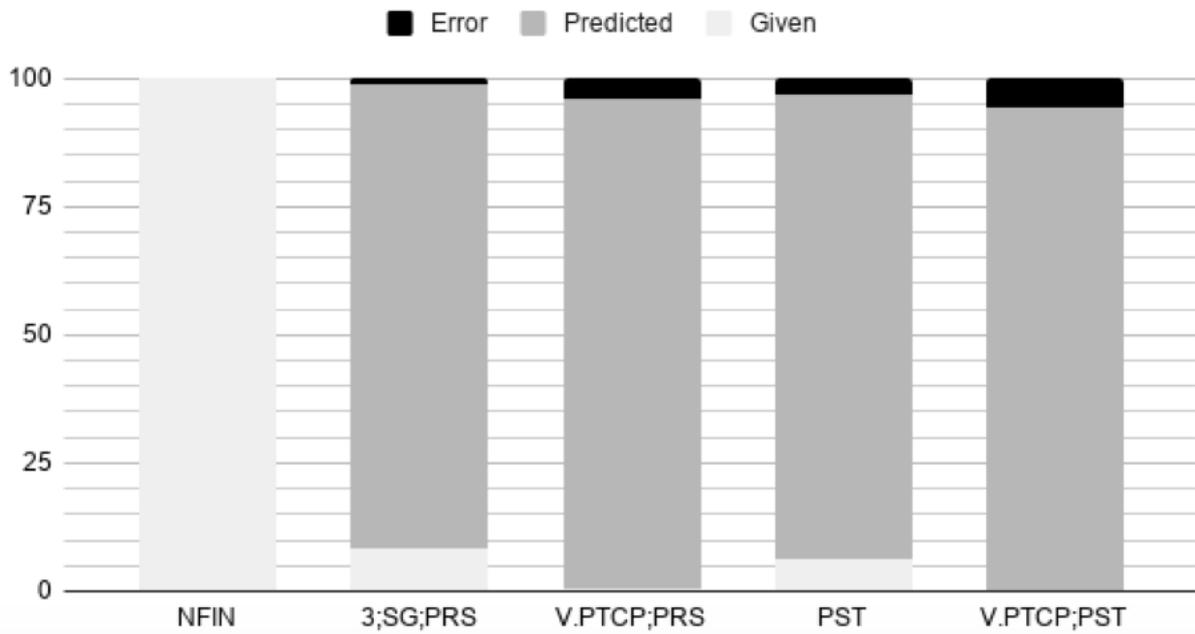
Figure 10: Coverage of English target cells after SSA chooses the first two sources.

English Verb Inter-Predictability

	NFIN	3;SG; PRS	V. PTCP; PRS	PST	V. PTCP; PST		Average Predictiveness
NFIN	1.00	0.97	0.95	0.89	0.89		0.94
3;SG;PRS	0.97	1.00	0.94	0.88	0.88		0.93
V.PTCP; PRS	0.91	0.91	1.00	0.91	0.91		0.93
PST		0.87	0.93	1.00	0.96		
V.PTCP; PST	0.87	0.88	0.94	0.96	1.00		0.93
Average Predictability		0.93	0.95	0.93	0.93		
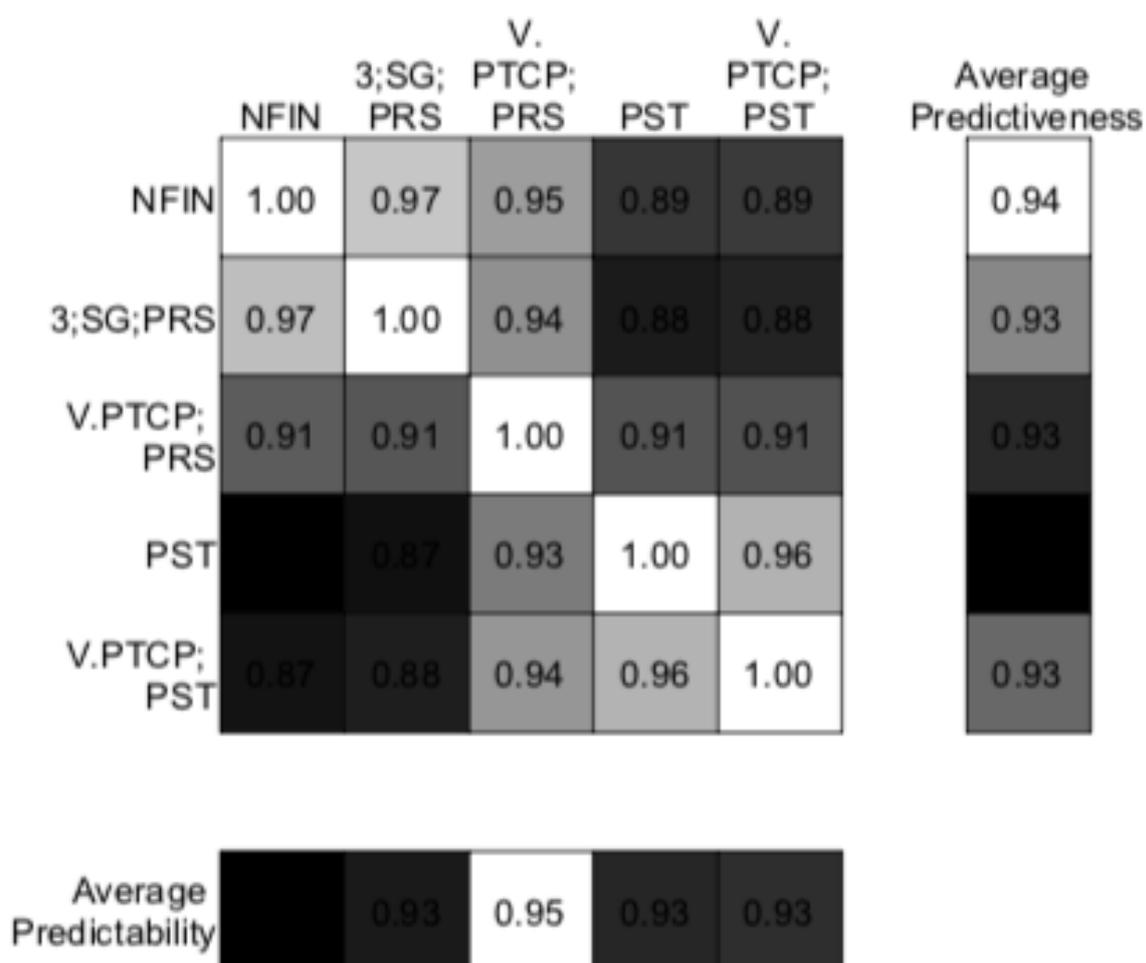
Figure 11: Inter-predictability heat map of English cells.
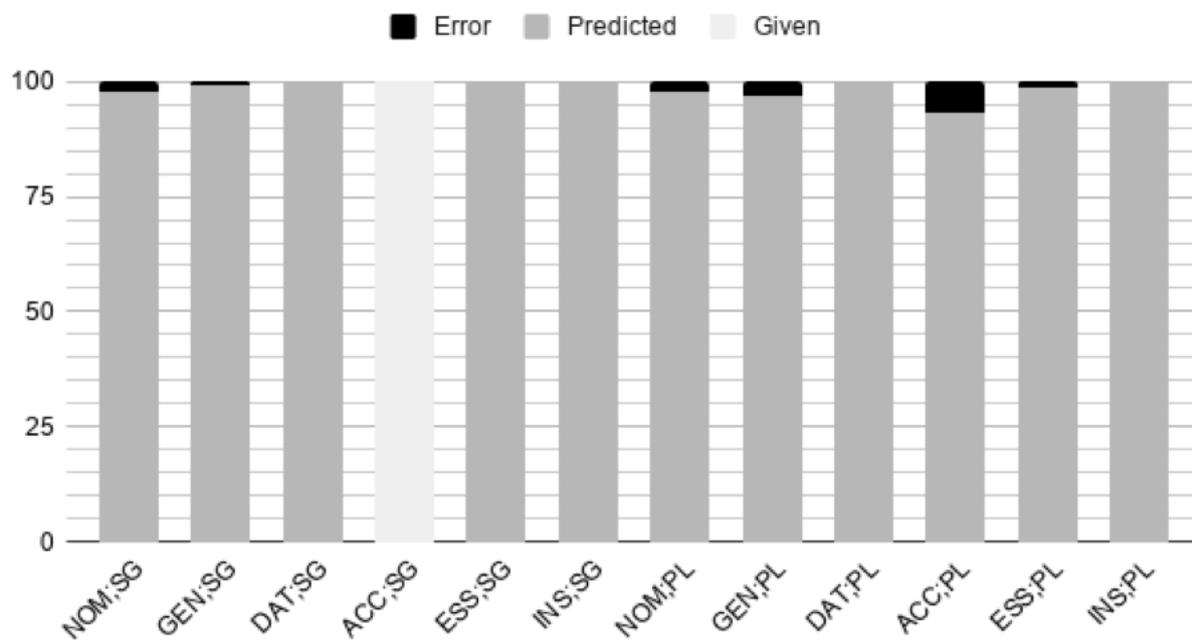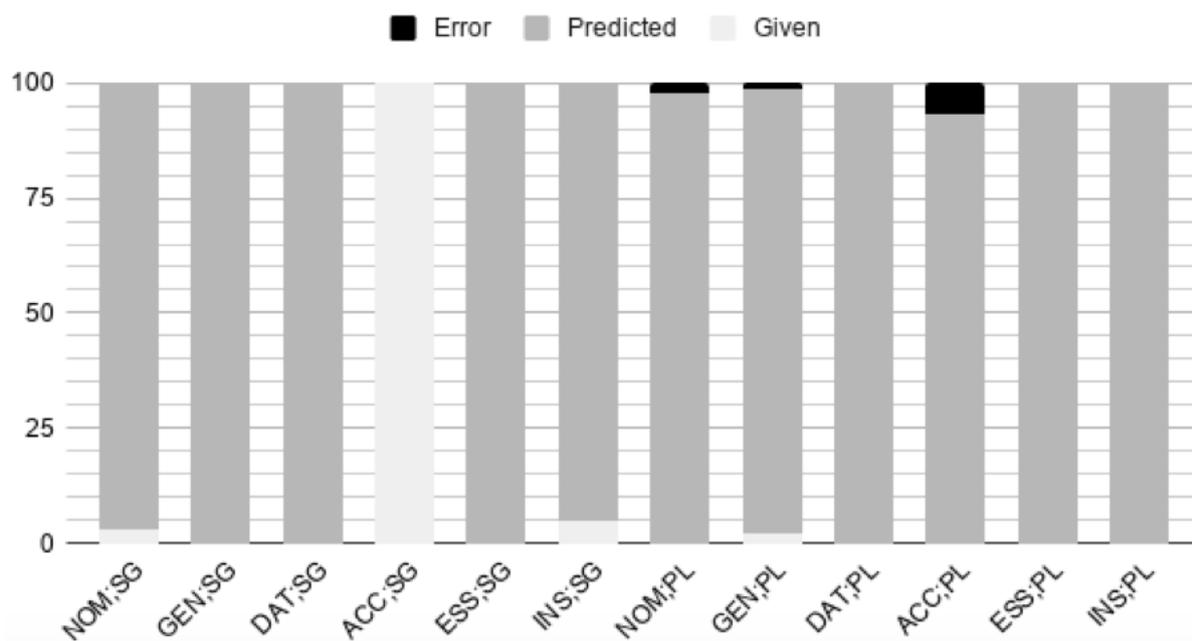
Figure 12: Coverage of Russian target cells after SSA chooses the first two sources.

Russian Noun Inter-Predictability

|	NOM; SG	GEN; SG	DAT; SG	ACC; SG	ESS; SG	INS; SG	NOM; PL	GEN; PL	DAT; PL	ACC; PL	ESS; PL	INS; PL		Average Predictiveness
NOM;SG	1.00	0.94	0.95	0.87	0.95	0.93	0.95	0.93	0.96	0.81	0.96	0.96		0.93
GEN;SG	0.91	1.00	0.96	0.82	0.96	0.94	0.95	0.92	0.97	0.80	0.97	0.97		0.93
DAT;SG	0.92	0.96	1.00	0.82	0.97	0.94	0.94	0.92	0.96	0.77	0.96	0.96		0.93
ACC;SG	0.97	0.95	0.95	1.00	0.95	0.93	0.95	0.93	0.97	0.93	0.96	0.96		0.96
ESS;SG	0.76	0.77	0.79	0.68	1.00	0.78	0.91	0.76	0.94	0.77	0.95	0.95		
INS;SG	0.93	0.97	0.97	0.82	0.97	1.00	0.94	0.92	0.96	0.80	0.97	0.97		0.94
NOM;PL	0.78	0.81	0.81	0.70	0.96	0.79	1.00	0.80	0.99	0.82	0.99	0.99		0.87
GEN;PL	0.90	0.90	0.90	0.77	0.92	0.90	0.92	1.00	0.95	0.78	0.95	0.95		0.90
DAT;PL	0.76	0.78	0.77	0.68	0.94	0.76	0.94	0.78	1.00	0.78	0.98	0.97		0.85
ACC;PL	0.79	0.82	0.81	0.80	0.94	0.80	0.95	0.79	0.96	1.00	0.96	0.96		0.88
ESS;PL	0.78	0.79	0.78	0.68	0.95	0.76	0.94	0.78	0.98	0.79	1.00	0.99		0.85
INS;PL	0.78	0.79	0.78		0.95	0.77	0.94	0.79	0.98	0.78	0.98	1.00		0.85
Average Predictability	0.86	0.87	0.87		0.96	0.86	0.94	0.86	0.97	0.82	0.97	0.97		

Figure 13: Inter-predictability heat map of Russian cells.
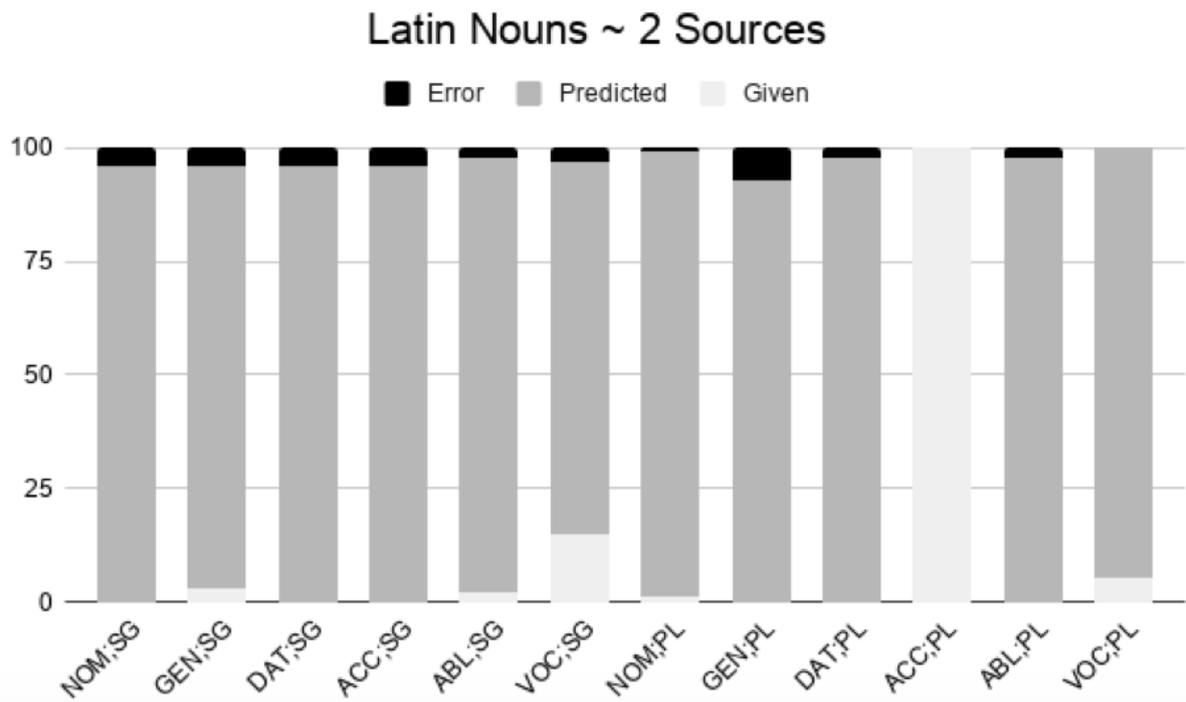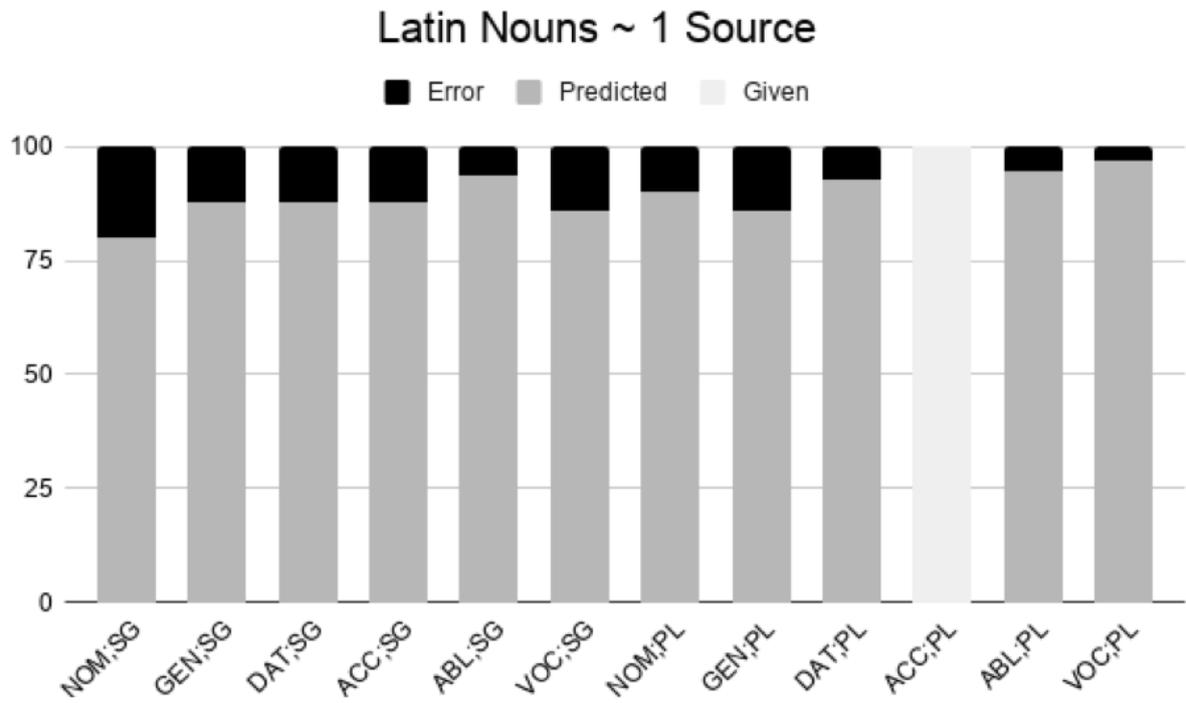
Figure 14: Coverage of Latin target cells after SSA chooses the first two sources.

Latin Noun Inter-Predictability

	NOM; SG	GEN; SG	DAT; SG	ACC; SG	ABL; SG	VOC; SG	NOM; PL	GEN; PL	DAT; PL	ACC; PL	ABL; PL	VOC; PL	Average Predictiveness
NOM; SG	1.00	0.81	0.80	0.79	0.81	0.87	0.81	0.80	0.80	0.78	0.81	0.82	0.82
GEN; SG	0.75	1.00	0.93	0.87	0.92	0.77	0.82	0.89	0.93	0.82	0.93	0.84	0.87
DAT; SG	0.72	0.89	1.00	0.85	0.90	0.73	0.80	0.86	0.92	0.79	0.91	0.82	0.85
ACC; SG	0.70	0.85	0.85	1.00	0.87	0.71	0.78	0.83	0.87	0.78	0.88	0.82	0.83
ABL; SG	0.77	0.93	0.93	0.88	1.00	0.77	0.84	0.89	0.92	0.84	0.92	0.84	0.88
VOC; SG	0.89	0.83	0.83	0.82	0.82	1.00	0.83	0.83	0.85	0.82	0.85	0.83	0.85
NOM; PL	0.83	0.89	0.89	0.85	0.90	0.82	1.00	0.85	0.91	0.90	0.90	0.93	0.89
GEN; PL	0.71	0.86	0.87	0.81	0.85	0.72	0.77	1.00	0.88	0.77	0.88	0.78	0.82
DAT; PL		0.66	0.68	0.63	0.69	0.58	0.60	0.63	1.00	0.64	0.90	0.65	
ACC; PL	0.80	0.88	0.89	0.87	0.91	0.82	0.86	0.85	0.90	1.00	0.90	0.95	0.88
ABL; PL	0.57	0.70	0.70	0.65	0.68	0.57	0.64	0.68	0.91	0.65	1.00	0.65	0.70
VOC; PL	0.81	0.90	0.91	0.89	0.90	0.81	0.93	0.87	0.91	0.94	0.91	1.00	0.90
Average Predictability		0.85	0.86	0.83	0.85	0.76	0.81	0.83	0.90	0.81	0.90	0.83	

Figure 15: Inter-predictability heat map of Latin cells.

Hungarian Nouns ~ 1 Source

■ Error ▨ Predicted □ Given
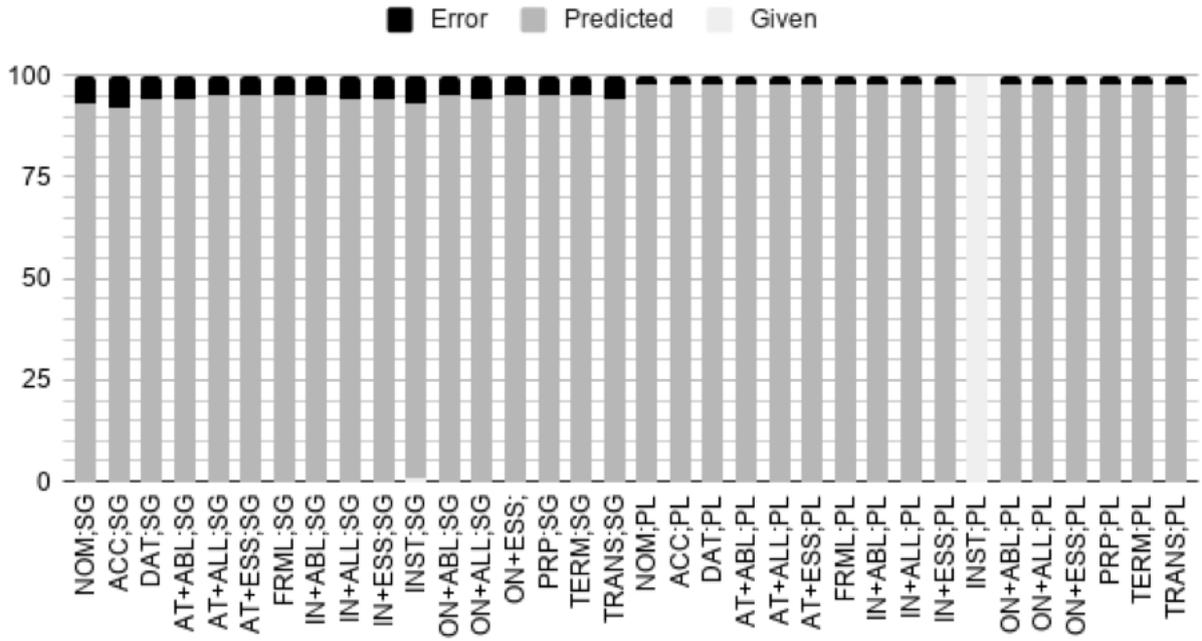
Hungarian Nouns ~ 2 Sources

■ Error ▨ Predicted □ Given
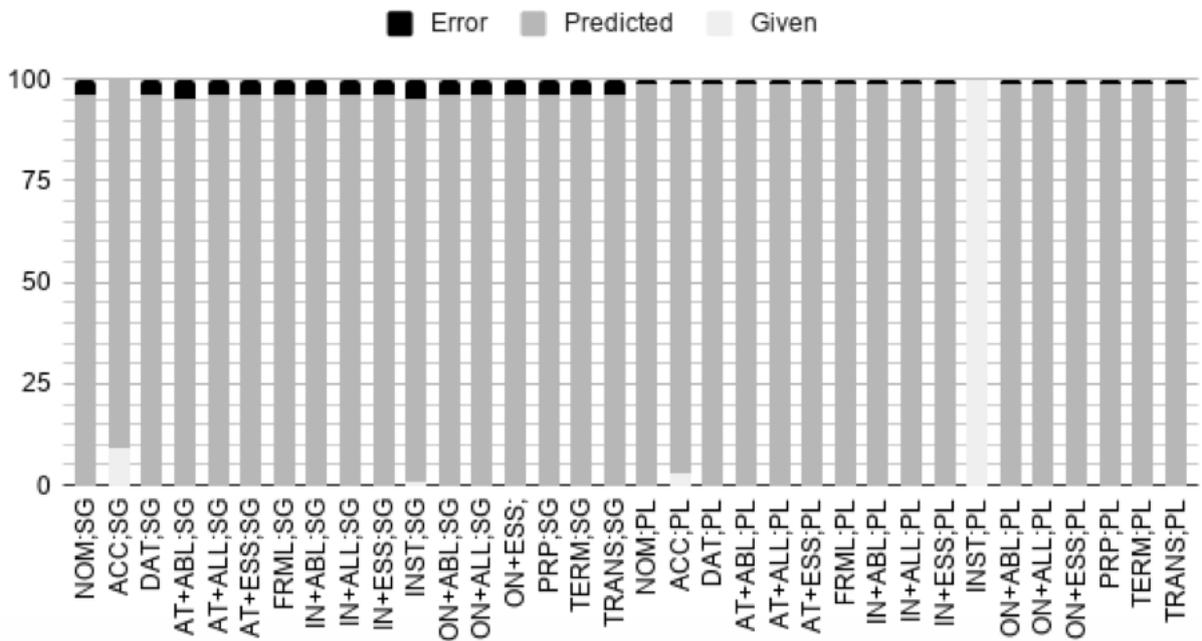
Figure 16: Coverage of Hungarian target cells after SSA chooses the first two sources.

Figure 17: Inter-predictability heat map of Hungarian cells.
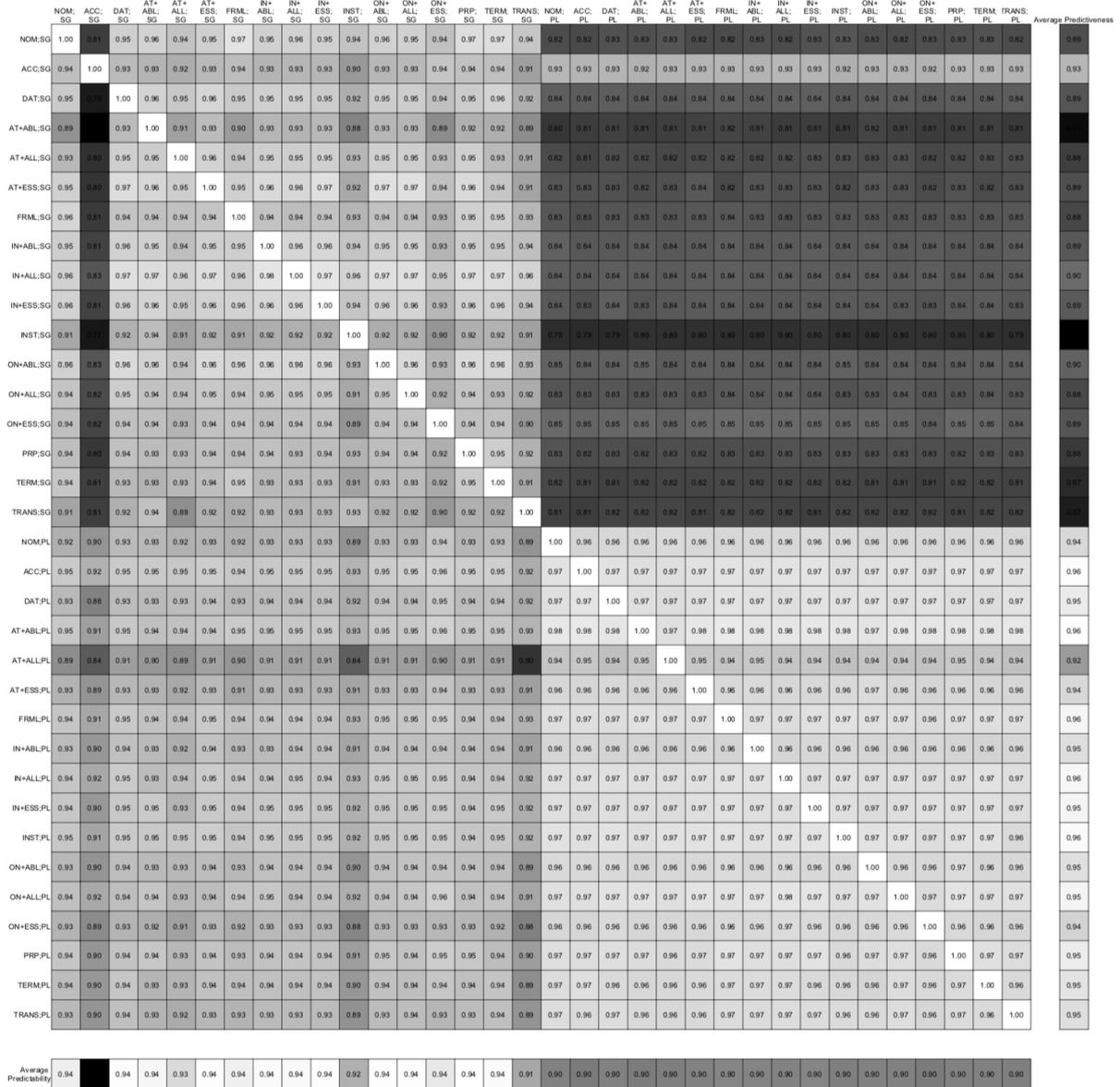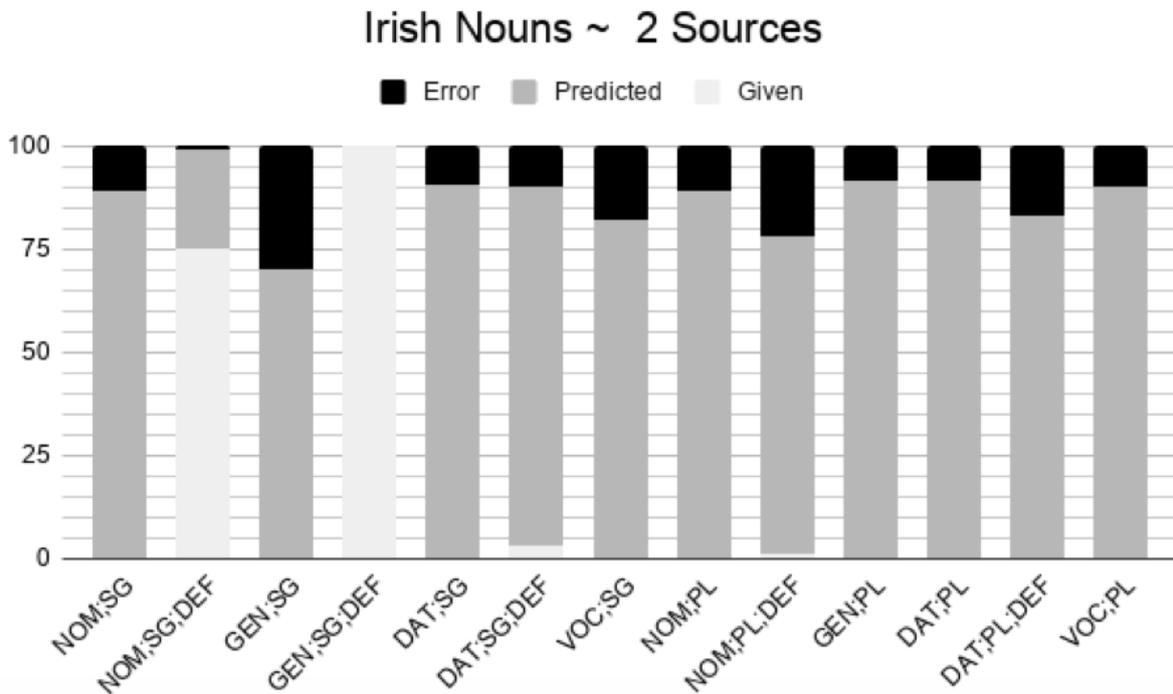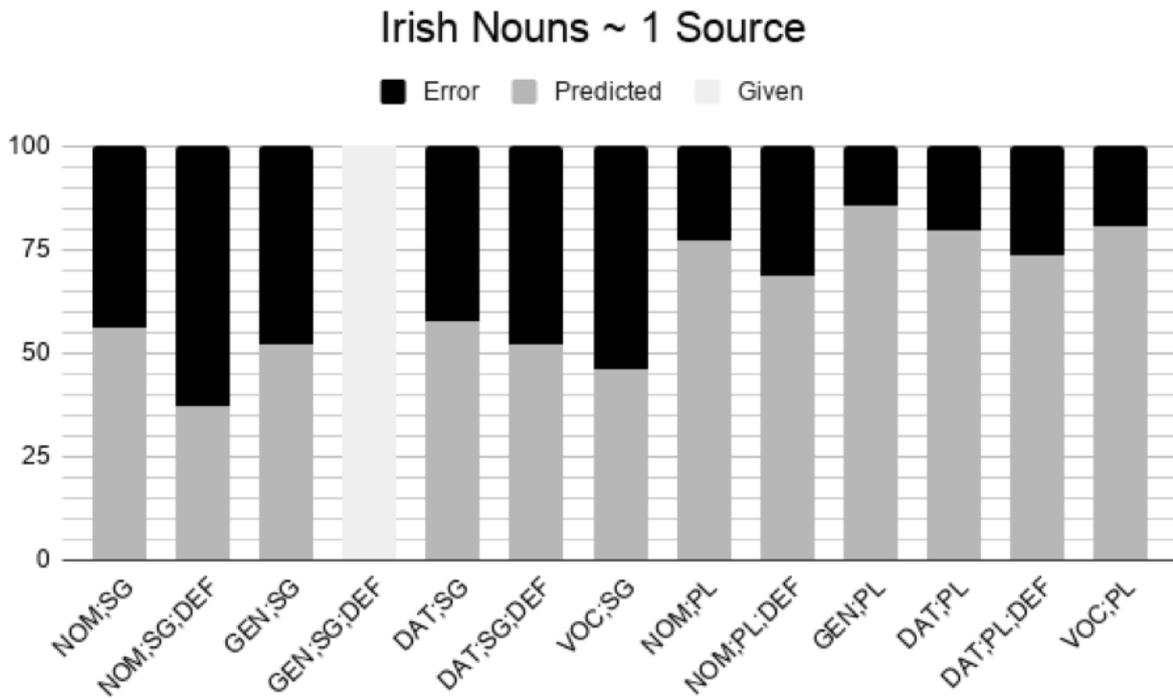
Figure 18: Coverage of Irish target cells after SSA chooses the first two sources.

Irish Noun Inter-Predictability

|	NOM;SG	NOM;SG;DEF	GEN;SG	GEN;SG;DEF	DAT;SG	DAT;SG;DEF	VOC;SG	NOM;PL	NOM;PL;DEF	GEN;PL	DAT;PL	DAT;PL;DEF	VOC;PL	Average Predictiveness
NOM;SG	1.00	0.36	0.26		0.50	0.44	0.42	0.16	0.17	0.17	0.16	0.17	0.21	0.32
NOM;SG;DEF	0.51	1.00	0.28	0.16	0.50	0.47	0.40	0.18	0.18	0.22	0.18	0.17	0.22	0.34
GEN;SG	0.33	0.21	1.00		0.32	0.26	0.29	0.21	0.21	0.20	0.22	0.19	0.19	
GEN;SG;DEF	0.29	0.17	0.25	1.00	0.28	0.24	0.27	0.43	0.44	0.44	0.43	0.46	0.45	0.40
DAT;SG	0.54	0.38	0.34	0.14	1.00	0.51	0.44	0.20	0.21	0.20	0.20	0.21	0.23	0.35
DAT;SG;DEF	0.59	0.44	0.31	0.17	0.58	1.00	0.45	0.21	0.22	0.23	0.20	0.21	0.23	0.37
VOC;SG	0.39	0.26	0.29		0.37	0.35	1.00	0.21	0.19	0.16	0.21	0.17	0.18	0.30
NOM;PL	0.32	0.18	0.29	0.41	0.28	0.28	0.30	1.00	0.48	0.40	0.49	0.49	0.45	0.41
NOM;PL;DEF	0.32	0.23	0.32	0.42	0.32	0.24	0.30	0.50	1.00	0.41	0.50	0.51	0.47	0.43
GEN;PL	0.27	0.20	0.27	0.40	0.26	0.27	0.23	0.41	0.38	1.00	0.41	0.38	0.43	0.38
DAT;PL	0.31	0.21	0.29	0.38	0.29	0.27	0.27	0.48	0.50	0.38	1.00	0.47	0.42	0.40
DAT;PL;DEF	0.28	0.22	0.28	0.41	0.27	0.27	0.28	0.49	0.53	0.42	0.48	1.00	0.45	0.41
VOC;PL	0.22	0.16	0.17	0.29	0.23	0.21	0.19	0.33	0.32	0.35	0.34	0.33	1.00	0.32
Average Predictability	0.41		0.33	0.32	0.40	0.37	0.37	0.37	0.37	0.35	0.37	0.37	0.38	

Figure 19: Inter-predictability heat map of Irish cells.