The Tense Debate

Cognitive Modelling English Past Tense Inflection with Encoder-Decoder Neural Models

Andrea Roijen
5726042
a.roijen@students.uu.nl

First Supervisor
Prof. Dr. A. Gatt

Second Supervisor
Dr. T. Deoskar

Master of Science Artificial Intelligence
Department of Information and Computing Sciences
Utrecht University
November 2024

Abstract

A longstanding debate surrounds modelling English past tense inflection. In this thesis, we investigated neural models, specifically Encoder-Decoders, as cognitive models of English past tense inflection. While recent studies showed that Encoder-Decoders achieve strong improvements over the initial connectionist model of Rumelhart and McClelland (1986), it has also been reported that they still failed to accurately capture human speaker inflections of novel forms. Our results reveal that data representativeness and model configuration choices influence model performance on real and nonce verbs. Importantly, we found improved correlations with human nonce verb inflections when the problem of overfitting on training data was mitigated, for instance, by using fewer training epochs. However, this also resulted in lower accuracies on real irregular verbs. A key finding is that this problem could be overcome by augmenting the training data with token frequency. This led to near-perfect performance on training verbs, including high accuracies on the irregular class, while obtaining almost equally strong correlations with human data. This highlights the relevance of token frequency, challenging previous assumptions. Additionally, we investigated a multi-task training setup, wherein the model also classifies verbs as regular or irregular. This task aligns with the dual-route view of Pinker and Prince (1988). However, this setup led to similar or slightly worse performance, leaving the cognitive validity of the discrete distinction between regular and irregular verbs open to further investigation. We emphasise the value of future research on using neural models to investigate cognitive processes such as morphological inflection.

Acknowledgements

I would like to express my sincere gratitude to my supervisors Albert Gatt and Tejaswini Deoskar. Thank you for your guidance and feedback, from which I have learned so much. I especially appreciated our regular brainstorming sessions during our meetings, which I found very inspiring and motivating. I am also grateful for your patience and understanding, without which this project would not have been possible.

I also wish to thank my family and Arjan for their unwavering support, love, and humour, which always help me through anything.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Morphology, the study of word formation, offers insight into the underlying mechanisms of our language abilities and therefore human cognition. A prominent area of focus in the study of morphology is the past tense inflection of verbs in English. It is an intriguing example because of its complex system that challenges our understanding of how the brain handles regular patterns like adding *-ed* (e.g. *walk–walked*) as well as less frequent exceptions to this rule (irregular verbs such as *do–did* and *see–saw*). Different approaches and models of the English past tense inflection have been proposed throughout the past decades, with the goal of obtaining a better understanding of the cognitive processes that underlie this case of morphological inflection.

With the developments of neural networks in the 1980s, Rumelhart and McClelland (1986) (R&M) introduced a connectionist model for the English past tense inflection. This model consisted of one input and one output layer of neural units, mapping both regular and irregular verbs to past tense forms. Such a model learns distributed representations based on patterns that the model finds in a dataset of verbs. This way of modelling morphological inflection therefore suggests that it is not necessary to explicitly define a set of rules, but that the language mechanisms can be characterised by rules that it finds as patterns in a dataset.

At the time, this rather controversial approach of R&M led to an extensive rebuttal by Pinker and Prince (1988) (P&P). According to P&P, defining explicit rules is essential in modelling the English past tense inflection. Moreover, they argued that the past tense inflection is governed by two systems. The regular inflection of verbs relies on a rule-based system, requiring only the memorisation of the verb stem and the regular transformation. The other system is responsible for the irregular inflection, relying on memory (and sometimes through gradient analogical processing, Prasada and Pinker, 1993). P&P showed that this *dual-route* view better accounts for human behaviour concerning irregular and regular inflections, and they emphasized the weak empirical performance of the R&M model. P&P ascribed the poor performance of R&M's model to the lack of a symbolic processing component representing explicitly defined rules for the regular inflection. Therefore, they claimed that this issue would arise with any connectionist model. This rebuttal by P&P had a significant influence on the field of linguistics, causing a broad dismissal of connectionist modelling.

Now, over 35 years later, incredible progress has been made in the field of Artificial Intelligence with deep learning developments. This led to far more sophisticated neural network models than those in 1986. One important development in the field of natural language processing (NLP) was Elman's (1990) introduction of Recurrent Neural Networks (RNNs). Building on these vanilla RNNs, further advances have been made, including Gated Recurrent Units (GRU, Cho et al., 2014) and Long Short-Term Memory (LSTM, Hochreiter and Schmidhuber, 1997). In addition, the introduction of sequence-to-sequence frameworks and attention mechanisms (Bahdanau et al., 2014; Sutskever et al., 2014; Vaswani, 2017) marked significant progress in the field of NLP. These developments offer advantages such as sequential processing and variable length inputs, which are especially beneficial for language modelling.

This progress inspired studies to revisit the English past tense modelling debate. Kirov and Cotterell (2018) (K&C) implemented a character-level Encoder-Decoder (ED) architecture to inflect English verbs to their past tense forms. Their model achieved near-perfect accuracy on all verbs encountered during the training phase, as well as on regular verbs from a held-out test set. The vast majority of the errors the model made consisted of over-regularisations of held-out irregular verbs. These errors, however, align with the human inclination to treat new verbs as regular forms (Albright and Hayes, 2003; henceforth A&H). Compared to the relatively poor empirical performance and less plausible errors of the R&M model, the K&C model showed a large improvement. Based on these results, K&C concluded that it is possible to model past tense inflections without defining explicit rules in a model and emphasise the importance of focusing on whether the model mimics human behaviour around novel input. This reveals whether the model generalises to a system of inflection that is similar to that of human speakers.

However, comparing a model's inflections of real verbs from a held-out set poses the challenge that we compare model inflections on verbs that it has not seen during training, while it is uncertain how these verbs would exactly have been inflected by human speakers at the first time encountering them. Instead, a golden test for evaluating a model's generalisations to novel inputs is considered the *Wug Test* (Berko, 1958). During this test, human speakers are prompted to produce inflections of nonce words, plausible word forms that do not really exist, such as *spling* (a nonce verb to be inflected to a past tense form such as *splinged* or *splang*). A&H conducted such a Wug Test for English past tense inflection, by eliciting past tense forms of 58 monosyllabic nonce verbs from 42 human speakers. The forms produced by human speakers on the Wug Test reveal which inflection patterns are preferred to apply to unseen forms, reflecting their productivity. Comparing human speaker nonce verb inflections to those of a model allows us to appropriately evaluate how human-like a model's generalisations are.

Following K&C in implementing the Wug Test to evaluate the model's generalisations, Corkery et al. (2019) presented a more comprehensive comparison between the nonce verb inflections of a similar ED model and human speakers from the A&H experiment. They found that the model and human subjects showed a similar general tendency to inflect novel verbs regularly. Also, if the nonce verb resembles a cohort of phonetically similar irregular verbs, both human speakers and the model were more likely to produce an irregular past tense form. However, the model inflected these nonce verbs more often irregularly than the human subjects. For example, the model inflected *spling* almost always to *splung*, following the irregular pattern of *sting–stung*, while the majority of human subjects inflected this nonce verb to *splinged*. This failure to precisely mimic human behaviour on novel input suggests that these ED models cannot yet be considered appropriate models of human morphological processing.

1.1 Research Questions

The results of Corkery et al. (2019) and K&C demonstrate that, contrary to what has been widely believed for decades, neural models may have potential as a cognitive model of inflection. However, Corkery et al.'s (2019) more complete evaluation of their ED model's inflections indicates that there are tense inflection? still challenges faced, as they did not closely match those of human speakers on nonce verbs. The challenge remains to fully understand the potential of these neural models as a cognitive model of morphological inflection. The primary objective of this thesis is therefore to explore whether—and in which ways— ED models can be enhanced as a more accurate cognitive model of English past tense inflection. Addressing this inquiry will lead to a deeper understanding of the intricacies of morphological inflection within the context of human language acquisition. The central research question guiding this research is therefore as follows:

> *What is the potential of ED models as a cognitive model of English past*

To address this main question, this thesis focuses on four sub-questions. These do not only investigate whether an ED model's fit with human behaviour on both real and nonce verbs can be improved, but also provide insights into the role of different aspects in past tense inflection.

1.1.1 Token Frequency

In all of the above-mentioned studies, the training data of the model consists of a list where each verb appears with equal frequency. This approach is based on prior research suggesting that *type frequency*—the number of verb *types* following a given inflection pattern—is an important feature in past tense acquisition (e.g., Albright and Hayes, 2003; Baayen, 2009; Bybee, 1995; Pierrehumbert, 2001). On the other hand, in language learning, verbs are encountered in a specific token frequency distribution, where token frequency refers to the number of verb *tokens* that follow a given inflection pattern. First of all, it could therefore be argued that a realistic cognitive model should be able to encounter verbs in their token frequency distribution and still generalise to past tense forms in a similar way to humans.

Furthermore, as elaborated in Sections 2.3.2 and 3.4.1, the token frequency distribution of verbs may even carry useful information about the productivity of past tense inflection rules (Baayen, 2009), potentially leading to improved model performance (Ma & Gao, 2022). Investigating this aspect provides further insight into the role of token frequency in the acquisition of English past tense rules and its implications for cognitive modelling.

I. *Does training an ED model on token frequency instead of type frequency lead to more similar real and nonce verbs inflections to those of human speakers?*

We investigated this question by comparing two different kind of ED models, where one kind is trained on data where verbs are equally distributed (i.e. type frequency) and the the other kind of ED models is trained on dataset in which verb examples are added proportional to their token frequency distribution. These models are compared to each other on their performance on real verbs and their correlation with nonce verb inflections from human Wug Test data.

1.1.2 Dual-Route Multi-task Training

The second sub-question of this thesis explores the effect of an auxiliary task that predicts whether the verb in question is regular or irregular. This auxiliary task compels the model to categorise each verb into one of two distinct classes. According to the dual-route approach of P&P, this discrete differentiation between regular and irregular verbs is essential in modelling English past tense inflection. Adding such an auxiliary task to the main task of predicting verb inflections allows to investigate whether this discrete classification between regular and irregular verbs enhances the model's alignment with human behaviour. This, in turn, could provide more insight into the credibility of the dual-route approach proposed by P&P. The second sub-question is therefore as follows:

II. *Does training an ED model on an auxiliary task of distinguishing regular and irregular verbs lead to more similar inflections of real and nonce verbs to those of human speakers?*

We investigated this by comparing models that are only trained on the main task of predicting the correct verb inflection to those that are trained on the main task as well as the auxiliary task of predicting the verb class. Again, we compared these different models on their performance on real verbs as well as their fit with nonce verb inflections from human Wug Test data.

1.1.3 Model Configuration Choices

While the previous two sub-questions touch on aspects that may also shed light on the underlying cognitive mechanisms of inflection, the third sub-question focuses on exploring the effect of different model configuration choices on the performance of the ED models. In the studies of Corkery et al. (2019) and K&C, there is no report of exploring different options for the data, training and architecture of their ED models for this specific task. Their settings led to a near-perfect performance on real verbs. However, a cognitive model should also capture human behaviour on novel inputs. Moreover, Corkery et al. (2019) results suggest that their model may overfit on real verbs from the training set. It is therefore relevant to explore the effect of different model configurations on both real and nonce verb inflections. Moreover, how performance is affected may also differ per model type: with or without token frequency represented in the data and with or without multi-task training. Therefore, our third sub-question is as follows:

III. *Which data and model configuration choices affect an ED model's ability to predict more similar inflections of real and nonce verbs to those of human speakers*

To answer this question, we conducted one experiment that explores different versions of our ED models (Experiment 1, Section 5.1). We subsequently compared model performance on real and nonce verbs when including a set of verbs that we have found to be missing in the dataset of previous studies (A&H, Corkery et al., 2019, K&C), with or without allowing verbs to have more than one correct inflection, the use of an early stopping mechanism with different settings, and other hyperparameters that influence a model's learning capacities but also the potential tendency to overfit on real verbs: number of layers, layer sizes, dropout, batch size, learning rate. The aim of this experiment is (1) to provide insight into the impact of different data, model, and training settings on modelling English past tense inflection; and (2) to finally select models that most accurately capture human behaviour on novel inputs.

1.1.4 Impact of the Verb Distribution

Finally, the fourth and final sub-question focuses on investigating the influence of the distribution of real verbs over a training and development set. Although in Corkery et al. (2019) all real verbs were utilised as training data, we follow K&C in splitting the set of real verbs into multiple sets. Our main reason for this is to prevent overfitting by tracking development set accuracy during training. Although it is a valid and common approach in machine learning to use a random division and training-development split of the data, we argue that it should be taken into account with this type of verb data whether different splits lead to different results. When it comes to the task of past tense inflection, selecting 20% of the verbs as development set could potentially affect representativeness of both the training and the development set, which is mainly with regard to irregular verbs (further elaborated in Section 5.2). This could also influence a model's fit with human behaviour. This is not taken into account by K&C. Our fourth and final sub-question is therefore as follows:

IV. *Does the distribution of real verbs over a training and development set influence an ED model's fit with human behaviour on real and nonce verb inflections?*

To answer this question, we conducted a second experiment, where we evaluated and compared our ED models' performance using different training-development splits. (Experiment 2, Section 5.2).

Finally, the results of the two experiments are aggregated to gain insight into the overall performance of the models trained on type versus token frequency, and single versus multi-task training.

1.2 Outline

In order to find answers to the posed research questions, the outline of this thesis is as follows. In the Theoretical Background in Chapter 2, we explain key concepts and theories that are relevant to this thesis. In Chapter 3, we offer a brief summary of the related literature that preceded this thesis. Chapter 4 provides an overview of the methods. Chapter 5 presents both Experiment 1 (Section 5.1) and Experiment 2 (Section 5.2). In Chapter 6 we present the final overall results, aggregating the results from both Experiment 1 and 2. We discuss the implications of these results in light of the research questions in the Discussion in Chapter 7, along with suggestions for future research. Finally, in Chapter 8 we draw the main conclusions from this thesis.

Chapter 2

Theoretical Background of Modelling Past Tense Inflection

2.1 English Past Tense Inflection

Since it is essential to understand the case of inflection before modelling it, this section offers a brief overview of English past tense inflection. In English past tense inflection, we generally make a distinction between regular and irregular verbs. These two classes are discussed in the section below.

2.1.1 Regular Verbs

The vast majority of English verbs belong to the regular class, approximately 97%. They are inflected by taking the stem of the verb and then adding the suffix *-ed*, e.g., *walk–walked*. The pronunciation of this suffix depends on the phonetic context. The suffix *-ed* can be pronounced as [-əd] or [-ɪd], if the verb stem ends in the sounds /t/ or /d/, such as the verb ['stɑːrtɪd] (*started*). Otherwise, if the verb base ends in a voiceless consonant (e.g., /p/, /k/, /f/, /s/, /ʃ/) a /t/ is pronounced, for example in [læft] (*laughed*). Finally, an /d/ is pronounced if the verb stem ends in a voiced consonant (e.g., /b/, /v/, /z/, /n/, /ŋ/, /l/, /r/) or a vowel, for instance [lʌvd] (*loved*) and [pleɪd] (*played*).

2.1.2 Irregular Verbs

Although the irregular class is much smaller than the regular class (the remaining 3% of the verbs), the top most frequently used verbs are irregular. Irregular verb inflection includes any other pattern than the regular inflection. Irregular inflection patterns can involve transformations like vowel changes (also known as ablauts, for instance, *sing–sang*), consonant changes (*make–made*), a combination of the two (*teach–taught*), or no change at all (*hit–hit*). Some irregular inflections are highly irregular such that the pattern is entirely unique (like the inflection *is–was*). However, there are also irregular verbs that can be considered semi-regular. These verbs follow similar inflection patterns and can therefore be generalized to some extent. For instance, the set {*wring–wrung, swing–swung, sting–stung, string–strung, ...*} includes verbs that all share the same irregular inflection pattern, where the vowel /ɪ/ changes to /ʌ/. Another clear example is the set of verbs where the vowel /ɪ/ changes to /æ/: { *sing–sang, spring–sprang, ring–rang, drink–drank, sink–sank, ...*}.

2.2 Theories of English Past tense Inflection

As mentioned in the introduction, it has been debated for decades how these English past tense rules are cognitively processed and should be modelled. The topic gained a lot of attention since the 1980s, when R&M introduced their connectionist model and P&P contradicted this type of modelling with their dual-route approach. Below, we give a brief description the dual-route and single-route views, as well as some views that fall outside or in between the two.

2.2.1 Dual-Route View

In the dual-route view, as proposed by P&P, the cognitive process underlying English past tense inflection is described by two distinct pathways. At encountering a verb, a speaker first determines whether the verb is regular or irregular, after which the appropriate route is activated.

One of these two routes accounts for rule-based transformations of verbs, specifically for the process of inflecting regular verbs. This is based on the fact that regular inflection consists of the deterministic morphological rule where the suffix *-ed* is added to the base form of the verb. Therefore, only the stem of regular verbs need to be memorised, together with the general transformation rule for all these verbs. According to P&P, a model of inflection therefore needs a component with explicitly defined rules to represent this part of the inflection process.

The other route describes a memory-based process of retrieving past tense forms of irregular verbs. The past tense forms of the irregular verbs are stored in a mental lexicon, and are retrieved during the past tense inflection process when needed. Instead of remembering the base of the verb and applying a general rule, the irregular verbs are stored as whole forms. For this route, the process relies on past exposure and memorising the irregular past tense forms, and sometimes generalising them by relying on gradient analogical processing (P&P; Prasada and Pinker, 1993).

2.2.2 Single-Route View

In contrast, the single-route view posits a unified mechanism for both regular and irregular verbs. This view suggests that both regular and irregular past tense forms are generated through a single pathway rather than of two distinct routes. Under this framework, inflection patterns are learned through a common process, such as relying on the analogical similarity between stored exemplars (Blything et al., 2018; Bybee, 1995; Seidenberg & McClelland, 1989)

A key example of a single-route model is also the connectionist model proposed by R&M. In such a model, a single network learns both regular and irregular inflections without relying on explicitly defined rules. Instead, this type of processing involves pattern recognition, which is done by adjusting the network weights based on the frequency of encountering inflection patterns, which can be considered implicit rule learning.

2.2.3 In-Between the Two Views

From a different perspective, modern neural networks could also be considered to not strictly adhere to either the single or dual-route view. Today, deep neural networks have a large capacity, which makes them able to obtain complex abstract representations. Although both regular and irregular verbs are still being processed within one network, the network representations could theoretically be complex enough to mimic both single-route aspects as well as dual-route aspects.

Another example of a model that can be considered to fall in between the two views is the minimal generalisation learner (MGL) of A&H. This non-neural model discovers multiple rules of English past tense inflection by inductive learning, to which it assigns confidence scores (i.e., stochastic rules). Such a model is in line with the single-route view, given that regular and irregular inflections are handled within one mechanism. It does not only define regular inflection rules, but also irregular inflection rules: think of the described semi-regular patterns like {*sing–sang, ring–rang, ...*}. Simultaneously, their findings align with P&P's perspective that past tense inflection is most accurately described by explicit rules, as their model outperformed a single-route analogical model relying on phonological similarity.

Finally, another modelling approach that falls in between the two approaches is neural models implementing a symbolic component. According to Marcus (1998, 2020) connectionist models have important limitations with regard to their potential to capture human cognition. Distributed representations might not be fully able to capture structured or rule-based cognitive processing. Therefore, Marcus advocates for an approach combining neural modelling and symbolic reasoning (or, more precisely, variables that facilitate generalisation), where the two work together in a complementary way rather than being mutually exclusive.

2.3 Morphological Productivity

For a model to accurately apply past tense inflection, it should capture the morphological productivity of inflection rules. Morphological productivity generally refers to the potential of a morphological pattern to be applied to novel forms (Bauer, 1983; Plag, 2003; Schultink, 1961). Accurately capturing the productivity of rules enables the model to generalise these rules to new words in a similar manner to human speakers, thus simulating the cognitive mechanisms underlying linguistic patterns observed in natural language use.

2.3.1 Wug Test as Test of Productivity

One way to investigate the productivity of morphological patterns is the Wug Test (Berko, 1958). As explained in the introduction, this test reveals how human speakers inflect unseen forms, which reflects productive patterns. In the original test of Berko (1958), participants were prompted to make plural inflections of nonce nouns: *one wug, two wug***s** (hence the name *wug* test).

A&H conducted such a Wug Test with the goal of testing how similar their model's verb inflections were to those of human speakers. In two experiments, they elicited from 42 human speakers 58 past tense inflections of monosyllabic nonce verbs. In the second experiment, the participants were also asked to rate suggested inflections of the nonce verbs. Based on the produced inflections of the participants, A&H computed for each nonce verb inflection a *production probability*. With these production probabilities, A&H evaluated the predicted confidence scores for nonce verb inflections by their MGL model (as briefly mentioned in Section 2.2.3). They did this by computing the correlation between their model's confidence scores and the human speaker production probabilities. This results in a single measure indicating the similarity between a model's and humans' generalisations to novel verbs.

2.3.2 Determiners of Morphological Productivity

2.3.2.1 Form Similarity and Type Frequency

Comparing model nonce verb inflections to those of humans, Corkery et al. (2019) found that their ED model failed at accurately capturing human behaviour. Therefore, it is interesting to consider what kind of linguistic information in the training input could contribute to generalising rules that lead to a stronger fit with human behaviour on nonce verbs.

For English past tense inflection, it is generally assumed that productivity mainly depends on the phonological form and the number of verb types belonging to the morphological class, i.e., type frequency (Bybee, 1995; Pierrehumbert, 2001; Skousen, 1989). Type frequency as a measure of productivity makes sense, given that around 97% of the English verb types belong to the regular class, the highly productive rule in English past tense inflection. Modelling evidence from A&H's study further showed that type frequency led to accurate predictions of novel verb inflections. Thus, the previous studies (Corkery et al., 2019; K&C) also trained their models based on a dataset in which each verb occurs once, hence type frequency.

2.3.2.2 Token Frequency

However, another option is to train the models on data that represents the token frequency of verbs. Such a decision would decrease the proportion of regular examples in the data; only $\sim 70\%$ of the verb tokens belong to the regular class, in contrast to $\sim 97\%$ regular verb types. As also visualised in Section 4.1.4 from our dataset, in the distribution of token frequency, the most common verbs appear significantly more frequently than the less common ones and they are often irregular. The regular class, on the other hand, contains relatively many verb types that occur less frequently. It might be seen as a potential drawback of token frequency that it represents less directly the high productivity of the regular inflection.

However, representing token frequency in the data may still be an interesting option when modelling past tense inflection. First of all, token frequency makes the model's input more similar to the input of human language learners. Humans acquire the past tense inflection by encountering verbs in their token frequency distribution, and not by hearing each verb equally often. It may be expected from a plausible cognitive model that it is able to learn from a similar input to that of human language learners. That is, the relevance of type and token frequency information should be learned by the model itself, like humans do. Thus, including token frequencies in the training input would contribute to a more representative cognitive model.

Moreover, token frequency may even offer information that is useful for the acquisition of the productivity of inflection rules. Baayen (2009) describes two measurements of morphological productivity in addition to *realized productivity* (based on type frequency): *expanding productivity* and *potential productivity*. Both these measurements are based on token frequency. To compute these measurements, Baayen (2009) uses the number of the words that occur once in the corpus per morphological class, hapax legomena. The more hapax legomena belong to a class, the higher the expanding and potential productivity of the class. These measurements are indicators of productivity of a morphological class, because a productive class is still expanding and applied to novel words Baayen (2009), which is indeed the case for the regular past tense rule. Following this rationale, we consider it interesting to investigate the effect of including token frequency on the model's ability to generalise in a human-like manner.

2.4 Encoder-Decoders as Models of Morphological Inflection

This final section of the Theoretical Background offers a brief description of character-level ED models, as implemented by this thesis, as well as by Corkery et al. (2019) and K&C. K&C implemented an architecture similar to that of Bahdanau et al. (2014) which Corkery et al. (2019) followed. EDs are relevant and successful models in the field of sequence-to-sequence tasks, especially in the field of NLP. The ED architecture is designed to process input sequences and generate corresponding output sequences, both of varying lengths. This makes these models highly suitable for natural language tasks such as machine translation (Bahdanau et al., 2014; Cho et al., 2014; Sutskever et al., 2014) but also morphological inflection (e.g., McCurdy et al., 2020; Corkery et al., 2019; Cotterell et al., 2016; K&C)

In the context of morphological inflection, the ED model processes inputs—in our case verb forms—as sequences of integers, where each integer represents a phonetic character. The encoder components process the input sequences by encoding them into fixed-length vectors. By passing the encoder states onto the decoder, the decoder will generate the output sequence, which is the past tense form of the verb.

In this architecture, both the encoder and decoder consist of an embedding layer. The embedding layers allow the model to build a more detailed representation of phonetic characters. Both the encoder and decoder contain two stacked LSTM layers. The input embeddings are fed into the first LSTM layer, which passes its output to the top LSTM layer. In the encoder, the LSTM layers are bidirectional, meaning that information is processed in both a forward and backward direction, instead of only forward (Schuster & Paliwal, 1997). The output vector of the encoder and the outputs of the decoder are passed onto the attention layer. From the attention layer outputs a context vector. Finally, a dense layer receives the output of the decoder together with the context vector. This dense layer predicts the next character in the output sequence. Both processing (encoder) and generating (decoder) verbs are done one character at a time. A schematic overview of the architecture is shown in Figure 2.1. Below, we also describe the main components of this ED architecture in more detail, along with their formal definitions.

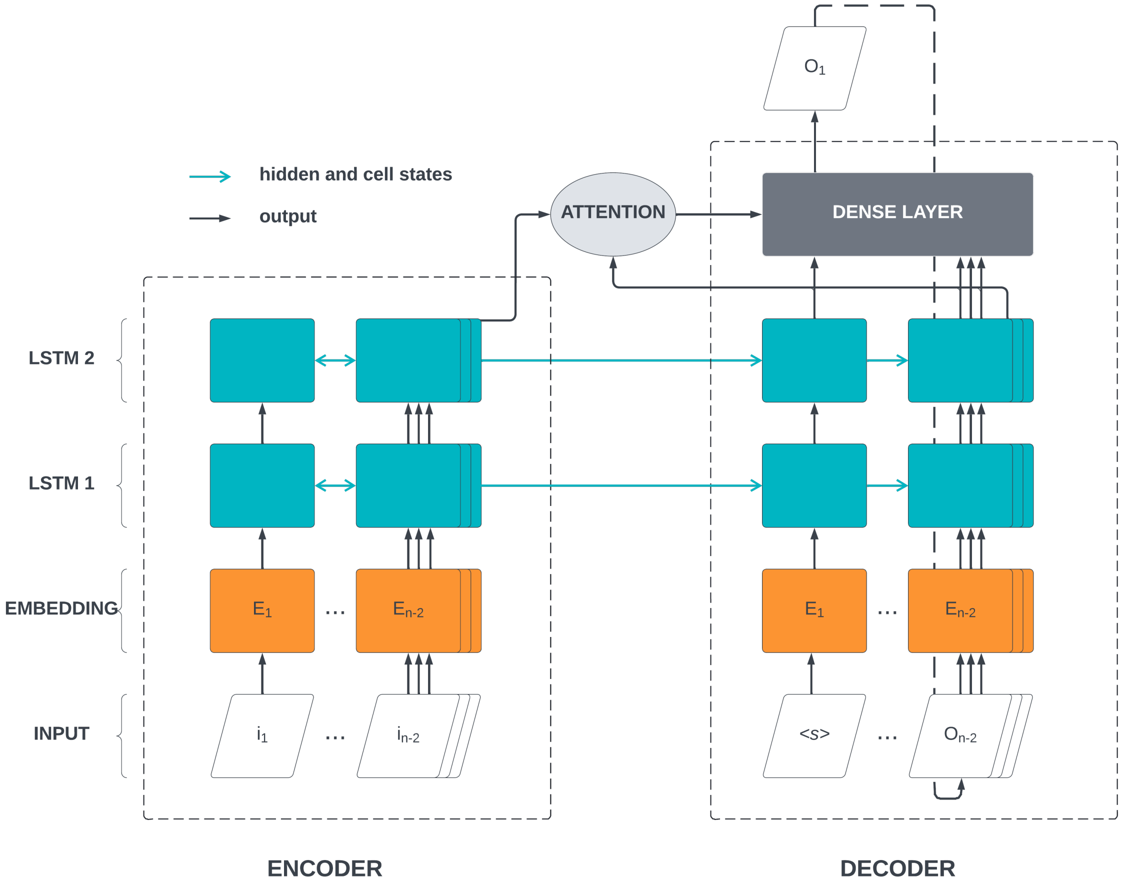


Figure 2.1: Schematic overview of the described encoder-decoder model.

2.4.1 Encoder

The input sequence $\mathbf{x}$ is fed into an embedding layer which converts each integer into a vector. Below a formal description of this is given, where x_1 is an integer representing one character of the phonetic form, and T is the length of this input sequence.

$$\mathbf{x} = (x_1, x_2, \ldots, x_T)$$
$$\mathbf{e}_i = \text{Embed}(x_i)$$

The first bidirectional LSTM layer processes the embedded input and generates an output based on its input at each time step. As the LSTM layer contains forward and backward hidden and cell states, $\mathbf{h}$ and $\mathbf{c}$, these are updated at each time step.

$$\overrightarrow{\mathbf{h}}_i, \overrightarrow{\mathbf{c}}_i = \text{LSTM}_{\text{fwd}}(\mathbf{e}_i, \overrightarrow{\mathbf{h}}_{i-1}, \overrightarrow{\mathbf{c}}_{i-1})$$
$$\overleftarrow{\mathbf{h}}_i, \overleftarrow{\mathbf{c}}_i = \text{LSTM}_{\text{bwd}}(\mathbf{e}_i, \overleftarrow{\mathbf{h}}_{i+1}, \overleftarrow{\mathbf{c}}_{i+1})$$

The forward and backward hidden and cell states together encapsulate information from both directions of the input sequence.

$$\mathbf{h}_i = [\overrightarrow{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i]$$

The output of the first LSTM layer is fed into the second bidirectional LSTM layer, resulting in the final encoder states and an output vector—the encoding—which abstractly represents relevant information from the input sequence.

$$\mathbf{h}'_i = \text{LSTM}_{\text{2nd}}(\mathbf{h}_i)$$

2.4.2 Decoder

Based on the encoder's final states, the decoder generates the past tense verb sequence. As mentioned, this is done character by character, as the decoder predicts the next character based on the previous character of the output sequence. The states of the decoder LSTM layers are initialised to the final hidden and cell states from the encoder, preserving information obtained by the encoder based on the input sequence.

$$\mathbf{h}_0^{\text{dec}} = \mathbf{h}^{\text{enc}}, \quad \mathbf{c}_0^{\text{dec}} = \mathbf{c}^{\text{enc}}$$

The first input the decoder receives is a start sign $< s >$. Next, each previously generated output character y_{i-1} is first passed onto the decoder's embedding layer, resulting in an embedding vector for that character.

$$\mathbf{e}_i^{\text{dec}} = \text{Embed}(y_{i-1})$$

As with the encoder, this embedding output is passed onto the first LSTM layer of the decoder. However, unidirectional LSTM layers are used in the decoder. Again, the hidden and cell states of the LSTM are updated at each time step.

$$\mathbf{h}_i^{\text{dec}}, \mathbf{c}_i^{\text{dec}} = \text{LSTM}^{\text{dec}}(\mathbf{e}_i^{\text{dec}}, \mathbf{h}_{i-1}^{\text{dec}}, \mathbf{c}_{i-1}^{\text{dec}})$$

The first decoder LSTM layer passes its output onto the second decoder LSTM layer.

$$\mathbf{h}_i^{\text{dec-2}} = \text{LSTM}_{2nd}^{\text{dec}}(\mathbf{h}_i^{\text{dec}})$$

2.4.3 Attention Layer

The output of both the encoder and decoder is passed onto an attention layer. The attention mechanism enables the model to concentrate on the most relevant parts of the encoder's output for the prediction of the next character. It does this by creating a context vector that indicates where information of the encoded input sequence is emphasised.

$$\mathbf{c}_i = \sum_{j=1}^{T} \alpha_{ij} \mathbf{h}_j^{\text{enc}}$$

Each annotation $\mathbf{h_j^{enc}}$ encapsulates information from the input, with most focus on the area of the j-th character of the input. The context vector is also computed based on an attention score α_{ij} (where i is the current time step of the decoder and j is the given time step of the annotations from the encoder output).

Depending on the type of attention that is used, the attention scores are computed differently. In the case of Luong attention (Luong et al., 2015), which is the mechanism we implemented in this thesis, the attention scores are based on the dot-product of the current decoder state and the encoder output states, after which Softmax is applied to normalise the weights.

$$\alpha_{ij} = \frac{\exp(\mathbf{h}_i^{\text{dec}\top} \cdot \mathbf{h}_j^{\text{enc}})}{\sum_{k=1}^{T} \exp(\mathbf{h}_i^{\text{dec}\top} \cdot \mathbf{h}_k^{\text{enc}})}$$

where $\mathbf{h}_i^{\text{dec}}$ is the hidden state of the decoder at time step i, $\mathbf{h}_j^{\text{enc}}$ is the hidden state of the encoder at time step j, and T is the length of the input sequence.

2.4.4 Dense Layer

Together with the decoder LSTM output $\mathbf{h}_i^{\text{dec-2}}$, the context vector $\mathbf{c_i}$ is fed into a dense layer at each timestep i. This dense layer predicts the next character of the output sequence y_i by using a Softmax activation. The process of predicting output characters is repeated until the stop sign is predict or when a defined maximum length of the output is reached.

$$P(y_i|\mathbf{h}_i^{\text{dec-2}}, \mathbf{c_i}) = \text{softmax}(W \cdot [\mathbf{h}_i^{\text{dec-2}}; \mathbf{c}_i] + b)$$

Where W and b are the weights and biases of the dense layer respectively.

In line with our second research question, concerning multi-task training, we experiment with an augmentation of this model where an auxiliary task is incorporated. This auxiliary task implies the addition of another dense layer. We describe the details of this layer in the Methods chapter (4.2.2).

Chapter 3

Related Literature on Modelling Past Tense Inflection

3.1 The Decades-Old Debate

Although the connectionist model of R&M was strongly countered by P&P with their dual-route approach, the work of R&M is still relevant and is part of a shift in cognitive modelling linguistic behaviour and how neural models could be used to do this. The results of this initial neural model of past tense inflection may not seem very convincing today, but considering the worse computational power and capacity compared to more contemporary neural models, it could be considered a decent achievement. Nonetheless, the shortcomings of the R&M model pointed out by P&P provided grounds for the wide rejection of neural networks in similar modelling tasks. Although P&P addressed many different theoretical and practical aspects of the R&M model to argue against neural modelling, central were the empirical limitations of the R&M model (such as an estimated 67% accuracy) that could not be overcome without a component in the model that represents the presence of explicit rule learning. Therefore, P&P argued that it was not just R&M's model, but neural models in general, that are not suited to model past tense inflection.

3.2 Neural Model Advancements since the Debate

Although the P&P arguments against connectionist modelling were highly influential in linguistics, many of these critiques may no longer hold up against today's neural networks. The improvements with deep neural networks in the past decades have been significant, making the capacities of neural models in the 1980s far behind that of contemporary models. This is not only thanks to the larger architecture and capacity of neural models, but also to developments such as the mentioned RNNs (Elman, 1990) and more sophisticated architectures—such as LSTMs (Hochreiter & Schmidhuber, 1997), GRUs (Cho et al., 2014), Transformers (Vaswani, 2017)—that are suitable for processing and generating linguistic data. As mentioned in the Theoretical Background (2), ED models and mechanisms like attention were introduced and further developed to address sequence-to-sequence tasks in NLP such as machine translation (Bahdanau et al., 2014; Luong et al., 2015; Sutskever et al., 2014). The main theoretical improvements of using EDs as models of inflection are that they can preserve the identity and order of the phonemes by processing them sequentially, learning embeddings for each phoneme, and allowing inputs and outputs of arbitrary length, whereas the R&M connectionist model lacked this capacity and flexibility.

3.3 Revisiting the Debate with Deep Neural Networks

Thanks to the above-described advancements, deep learning models have become a more popular approach in cognitive modelling. Specifically, with the impressive progress in NLP, modelling tasks such as morphological inflection have also been approached with deep neural models. Multiple studies have focused on reinflection tasks (transforming one inflected form into another inflection of that same form, Cotterell et al., 2016) and morphological paradigm completions (mapping lemma's to forms in a paradigm, e.g., *work–works/worked/working*), which are generalisations of the inflection task described by R&M (Ahlberg et al., 2015; Durrett & DeNero, 2013; Faruqui et al., 2015; Nicolai et al., 2015).

3.3.1 Introducing Encoder-Decoders as Models of Past Tense Inflection

As the next step, K&C implemented an ED model to revisit the R&M and P&P debate on modelling past tense inflection. They demonstrate that their model overcomes multiple limitations of R&M's connectionist model. They investigated to what extent these ED models are able to learn correct and human-like generalisations to inflect verb stems to their past tense forms on both training and test of real verbs.

To do this, they implemented an ED model, as globally described in Section 2.4, and trained their model on CELEX verb data (Baayen et al., 1995) containing around 4000 examples[1]. Their model obtained near-perfect accuracy on all verbs in the training set, as well as on regular verbs from a test set. This is indeed a large improvement on the estimated 67% performance of the R&M model. On the test set of irregular verbs, however, it still struggled to generalise. This makes sense, given that irregular verbs can contain little to no similar patterns to generalise. Nonetheless, the model still obtained almost 30% accuracy on these test irregulars, which implies that the model can generalise irregular forms to some extent. Moreover, its errors on the irregular test verbs are mainly due to over-regularisations. They argued that these over-regularisation errors are desirable since human speakers have the tendency to produce regular inflections of novel forms as well (A&H). The errors of the model contained no *blend* errors, where the regular and an irregular inflection are blended (like *see–sawed*) while this was another limitation of the R&M model[2]. In conclusion, K&C's study demonstrates that contemporary neural models overcome multiple limitations of the initial model by R&M, and they suggested focusing on whether the model behaves in human-like manner on novel inputs.

[1] Which is another improvement on the R&M model. At the time, less data was available, so R&M used only 506 examples, of which 98 irregular (Kucera & Francis, 1967), as was also pointed out by P&P.

[2] According to Pinker (1999), blend errors are only a human-like error in the case of children learning the irregular past tense as the present tense of that word.

3.3.2 Capturing Human Behaviour with Encoder-Decoder Models

K&C further investigated their model's fit with human behaviour using A&H's Wug Test data. As explained in Section 2.3.1, the Wug Test offers an appropriate test of generalisation. They found a moderate correlation between the human nonce verb inflections and their ED model's predictions. As a follow-up, Corkery et al. (2019) further investigated the predictions of a similar ED model on the Wug Test. Their more comprehensive comparison between nonce verb inflections of human speakers and those of their ED model showed that human behaviour is not appropriately being captured.

Corkery et al. (2019) implemented a completely similar ED model to that of K&C. Alternatively, Corkery et al. (2019) used all real verb data as training data and used only the Wug Test as final test of the model's generalisations. They could indeed mimic K&C's results in obtaining near-perfect accuracy on the training set of real verbs. However, they also found that models with different random initialisations exhibited varying levels of correlation strength with A&H human data of nonce verb inflections. They argued that the instability in inflection behaviour of their model can be compared to variation between human speakers.

For this reason, in a second experiment, they aggregated the results of multiple random initialisations of the same model, considering each simulation of the model an individual speaker. They mimicked the computation of A&H's production probabilities of human speakers by aggregating the nonce verb inflections of 50 simulations. Their results revealed that the ED model generally followed a similar pattern to that of the human speakers, applying most often the regular inflection on nonce verbs. On a more detailed level, however, they observed that the ED model is more likely to apply the irregular inflection, compared to human speakers. When a nonce verb resembles a cohort of phonetically similar *irregular* verbs (also referred to as an *Island of Reliability* by A&H), both the human subjects and the model were more likely to treat the nonce verb as irregular, compared to when this is not the case. However, this effect was stronger for the model than for the human speakers. The computed correlations between the aggregate model predictions and human data were moderate for the regular inflection and weak for the irregular inflection of nonce verbs. Moreover, correlations following from the rule-based model by A&H were stronger. In conclusion, taking a closer look at the ED model's generalisations, significant shortcomings of ED models were still found when it comes to mimicking human speakers on the Wug Test.

3.4 Improving Encoder-Decoders as Models of Inflection

It follows from Corkery et al. (2019) that the question remains whether ED models can be appropriate cognitive models of past tense inflection. In this section we discuss related literature in light of the four research questions (posed in Section 1.1). For each, we point out how the previously discussed and related literature leads to these questions.

3.4.1 Augmenting Data with Token Frequency of Verbs

As mentioned in the Theoretical Background (2), we consider integrating token frequency in the data more appropriate for a cognitive model of past tense inflection, as opposed to type frequency, because this is more similar to the linguistic input of humans. In addition to this, we explained that the token frequency distribution of verbs might even contain useful information with regard to the productivity of the inflection rules, which is essential in modelling morphological inflection (Baayen, 2009).

Although most studies do not include token frequency in any way, a recent study by Ma and Gao (2022) investigated the effect of different verb frequency and resampling methods in training Transformer models on English past tense inflection. One of their findings indicated that the best accuracy on irregular verbs was obtained when the model was trained on a dataset where irregular verbs occur more often and each regular verb once. Vice-versa, regular verb performance was more affected by type frequency. They concluded that performance is influenced by both token frequency and type frequency, but these influenced performance on irregular and regular verbs, respectively. According to Ma and Gao (2022), this indicates that their model obtained abstract representations that distinguish regular verbs from irregular verbs.

These results show that augmenting the data with token frequency can influence model performance. Therefore, we investigated the effect of token frequency on the performance of ED models like the ones used in Corkery et al. (2019) and K&C. Contrary to Ma and Gao (2022), however, we augmented our training data with the token frequency of *all* verbs (i.e., both regular and irregular), which mimics how verbs are encountered during human language learning. In the Methods chapter (4), we explain this in more detail.

3.4.2 Multi-task Training the Dual-Route

Our second research question relates to Ma and Gao (2022), who stated that their Transformer model seemed to learn an abstract distinction between regular and irregular verbs. Contrary to the classic view that neural models oppose the dual-route approach, deep learning models could theoretically learn abstract representations that align with the dual-route perspective, if such representations help the model identify inflection patterns in the data. K&C also pointed out that they wanted to bypass the original question from the P&P and R&M debate "whether or not neural models learn and use "rules." From our perspective, any system that picks up systematic, predictable patterns in data may be referred to as rule-governed." (K&C, pp. 651–652). In fact, as mentioned, Marcus (1998, 2020) proposed representing explicit rules within a neural model. The integration of these—originally opposing—theoretical concepts in deep neural networks highlights the capacity of neural models to function as a means to investigate ongoing questions in psycholinguistics and cognition, such as past tense inflection.

To our knowledge, integrating a dual-route approach into a deep learning model has not been explored. One way to investigate the effect of integrating a dual-route approach within neural models is to facilitate the model in making a binary distinction between regular and irregular verbs. As mentioned in the introduction and further explained in the Methods chapter (4), we added an auxiliary task where the ED model predicts whether verbs are regular or irregular, in addition to the main task of predicting the past tense form. Investigating the effect of adding such an auxiliary task reveals whether or not the imposed emphasis on a binary distinction between regular and irregular improves the model's fit with human data. This, in turn, gives more insight into the plausibility of the dual-route view, which argues that a discrete distinction between regular and irregular verbs is in line with how humans process English past tense inflection.

3.4.3 Model Configuration Choices

For the third sub-question of this thesis, we investigated the effect of different model configuration choices for EDs as models of inflection. We consider this relevant, as it is known that aspects such as data, duration of training, architecture and the complexity of a neural model can heavily influence performance.

An initial consideration could perhaps be to explore more state-of-the-art (SOTA) neural architectures, like Beser (2021) and Ma and Gao (2022), who implemented transformer models instead of the ED architectures with LSTM-based mechanisms as implemented by Corkery et al. (2019) and K&C. However, the goal of this thesis is to follow up on the studies of Corkery et al. (2019) and K&C and to investigate whether performance improvement can be found as a result of the discussed research questions so that we improve our ability to model and understand past tense inflection. Therefore, SOTA performance is less of interest here, and we followed the previous studies of Corkery et al. (2019) and K&C by implementing a globally similar model.

However, Corkery et al. (2019) and K&C do not report to have explored alternative settings to the ones they used. To begin with, Corkery et al. (2019) adopted both A&H's and K&C's verb data, noting that K&C's set includes fewer examples. Nonetheless, even the A&H data excludes a significant set of mainly irregular verbs with a high token frequency. We provide more information about this omitted verb set in the Methods (Section 4.1.2.2). We investigated the effect of this omission of verbs and experimented with making the data more complete. We also investigated the effect of including and excluding examples of verbs if they have more than one correct inflection.

Furthermore, both Corkery et al. (2019) and K&C trained their models until near-perfect performance on training verbs using a fixed number of training epochs (100). However, Corkery et al. (2019) noted that the beam probabilities were heavily skewed, meaning that the top prediction in the beam has a very high probability, while the other predictions have very low probabilities. With fewer training epochs, Corkery et al. (2019) attempted to achieve more stable beam rankings and potentially better correlations. After every 10 training epochs, they computed the correlation between their model's nonce verb inflections and those of human speakers. Surprisingly, the highest correlation was obtained after just 10 training epochs with indeed more stable beam rankings. However, due to the low number of training epochs, the accuracy on real verbs was much worse. This was especially true for the set of irregular verbs, on which only 6.5% training accuracy was obtained instead of near-perfect accuracy. According to Corkery et al. (2019), these results imply a disparity in capturing human behaviour accurately on real verbs versus on nonce verbs.

From these results, one could suspect that a too-high number of training epochs leads to overfitting on real verbs and prevents human-like behaviour on the Wug Test. Additionally, the extent to which this is true may differ for each model (i.e., type frequency versus token frequency, single task versus multi-task training). As a potential solution to this problem, we explored the use of an early stopping mechanism to investigate how the number of training epochs influences performance on real and nonce verbs for each model.

Finally, we explored various model versions with different hyperparameter settings. K&C adopted hyperparameter settings from (Kann & Schütze, 2016), which Corkery et al. (2019) followed. However, Kann and Schütze (2016) focused on a morphological reinflection task. Since our exact task differs from theirs and hyperparameter settings can influence a model's learning capacity, complexity and the amount of regularisation, we experimented with various settings for the number of LSTM layers, layer sizes, dropout rate, batch size, and learning rate.

In Experiment 1 (Section 5.1), we elaborate on the procedure and results of the above-described model configuration experiments.

3.4.4 Distribution of Verbs

A final remark that we make on both Corkery et al. (2019) and K&C's studies is that they relied on either one random data split to divide their data into a training, development, and test set (K&C) or no split of real verb data at all (Corkery et al., 2019). Using all real verbs as training data eliminates the problem of omitting a significant subset of verbs from the learning data for the model. This is more representative of human linguistic input since the training data is as complete as possible. Omitting a set of verbs from the training set makes the training set less representative. Additionally, it is difficult to determine which set of irregular verbs would make a representative development or test set, given the remaining training data. In other words, which irregular verbs would optimally measure the model's ability to generalise? Verbs like *is–was* are nearly impossible to generalise to, but semi-regular verbs potentially make a more representative set of generalisations. However, it is uncertain which exact selection would be most representative.

On the other hand, dividing the real verbs into a training and development set allows for the use of preventive methods to mitigate overfitting on the training data. Even though we follow Corkery et al. (2019) in using the Wug Test as the ultimate final measure of generalisation, a development set allows for the use of an early stopping mechanism to monitor during training when the model begins to overfit on the training verbs.

As a solution to this dilemma, we followed K&C in making a random division of the data to be able to develop the model and look out for overfitting on real verbs in the training set. In a subsequent experiment (Experiment 2, Section 5.2), we investigated the effect of using different distributions of real verbs over the training and development set by training and evaluating our models on four alternative training-development set divisions.

Chapter 4

Methods

4.1 Data

4.1.1 Data Acquisition

4.1.1.1 Real Verbs

For this thesis, we adopted the set of CELEX verb lemmas used by A&H, K&C, and Corkery et al. (2019). We retrieved this set from Kirov (2023). The CELEX database is a lexical database for English, Dutch and German, containing detailed information such as lemma morphology, frequency, phonology, orthography and syntax (Baayen et al., 1995). From this CELEX database, A&H selected 4253 verbs, all with a lemma frequency of at least 10. For each lemma, the set contains its associated present and past tense forms and its lemma frequency. It is also indicated whether the past tense inflection is regular or irregular.

For the purpose of this thesis, we obtained the token frequency of the past tense forms from the CELEX database. Note that the lemma frequencies from the adopted A&H verb set represent the sum of the token frequencies of *all* forms of the lemma. The token frequency of the past tense forms, however, represents best how often the inflected form itself occurs.

Finally, we adopted K&C's American English phonetic representation of the verbs. Like A&H, Corkery et al. (2019), and K&C, we use phonetic representations instead of orthographic ones, as A&H's experiment was conducted in spoken form, and the mapping between phonology and orthography in English is not always direct. We use the American English phonetic transcriptions because this aligns with the language variety used in the A&H experiments, which were conducted with American English speakers.

4.1.1.2 Nonce Verbs

For this thesis, we follow Corkery et al. (2019) and K&C in adopting the Wug Test stimuli from A&H. This set consists of 58 nonce verbs. As for the real verb data, we relied on American English phonetic transcription of the nonce verbs (K&C) and retrieved this set from Kirov (2023). The Wug Test stimuli consist of present tense forms and multiple past tense forms for each nonce verb. These past tense forms always contain the regular past tense inflection and one or—occasionally—two possible irregular past tense inflections.

A&H based these past tense forms on their experimental production data. We used the past tense forms to classify the nonce verb inflections produced by our models. For the evaluation of our models, we compare our models' predictions to the production probabilities ($n = 44$) and inflection acceptability ratings from the A&H experiments ($n = 24$). This will be further explained in Section 4.4.2.

4.1.2 Data Preprocessing

4.1.2.1 CELEX Verb Selection

We took a subset of the original A&H real verb dataset consisting of verbs with a lemma frequency of 10 or higher. From this set, we excluded all verbs that have a past tense token frequency of 0. This eliminated 159 verbs from the A&H dataset. A few examples of these verbs are: *volleyed, dieted, gardened, gamed, caroused, mouldered, blubbered, clowned, fudged, quibbled, outran*.

4.1.2.2 Missing Verbs and Doubles

Furthermore, conducting a comprehensive examination, we found that a set of 25 verbs is actually missing from the A&H dataset. Table 4.1 shows 9 of the omitted verbs with the highest past tense form token frequencies. Though A&H do not give their reasoning behind the exclusion of these verbs, most of the verbs from Table 4.1 can be used as auxiliaries and are highly irregular past tense forms. Nevertheless, since the goal is to investigate the influence of token frequency and these forms are also highly frequent, we consider it essential to include them in the dataset. Moreover, because of their highly irregular form, these verbs could be expected to be memorised in human language learning, so it is valuable to see how the model handles such verbs.

Present tense	**Past tense**	**Token frequency past tense**	**Class**
is	was	97.174	irregular
have	had	22.393	irregular
are	were	8609	irregular
will	would	8609	irregular
can	could	8415	irregular
make	made	3572	irregular
may	might	1909	irregular
ought	ought	205	irregular
strike	struck	202	irregular

Table 4.1: Missing verbs from the A&H dataset.

Subsequently, we consolidated identical examples[1] from the dataset into a single example and summed their token frequencies. Further inspection of the dataset revealed that there are 61 past tense forms which share their present tense form with at least one other past tense form. This includes examples like {*ring–rang, wring–wrung, ring–ringed*}, as they have an identical (pronunciation of the) present tense form but a different past tense form.

[1] Two examples with the exact same present tense form and past tense form, meaning that they are duplicates.

These examples can differ in meaning and orthography, like *ring* and *wring*. It also includes examples with identical present tense forms, having two ways to be inflected, such as {*dive–dove*, *dive–dived*}. These 61 examples are referred to as *doubles* hereafter.

Previous studies have retained all past tense forms of these *doubles* in the dataset. This means that a model is confronted with the task of learning multiple inflections for the same input. Though it is most complete to include all forms in the dataset, it is worth noting that humans are confronted with these *doubles* in a semantic context that likely disambiguates these *doubles* (such as *ring* from *wring*). Nevertheless, like K&C, we consider the task of lexical disambiguation beyond the scope of this thesis.

To gain insight into the impact of including or excluding these missing verbs and *doubles*, this is the first element that is investigated in Experiment 1 (Section 5.1.2). These experimental results indicated that the set including all *doubles* and missing verbs is not only the most comprehensive and human-like version of the dataset, but also results in the best fit to human data. For this reason, all *doubles* and missing verbs were included in the dataset.

4.1.3 Character and Class Representation

As input, a character-level ED model needs float tensors of equal length. To do this, each phonetic character that occurs in the dataset was indexed with an integer. Each verb form was converted into a sequence of integers representing its phonetic representation. They were also zero-padded to the longest sequence in the dataset. Below an example of this is given.

pay	peɪ	[21 3 9 0 0 0 0 0 0 0 0 0 0 0 0]
paid	peɪd	[21 3 9 12 0 0 0 0 0 0 0 0 0 0 0]

For the auxiliary task, the verb class labels (*regular/irregular*) are indicated with binary labelling.

4.1.4 Data Augmentation with Token Frequency

The token frequency that we obtained is a count of how often the verbs occur in the corpora of the CELEX database. The total sum of token frequencies of all verbs in the dataset is 340.391. As can be seen from Figure 4.1, the token frequency distribution of verbs is heavily skewed. Normalising these raw token frequencies by taking their square root (Osborne, 2002) results in a less skewed distribution of the verb token frequencies. We show this distribution in Figure 4.2. As a result of the normalisation, the overall token count of all verbs was reduced to a total sum of 16.263 tokens. This smaller size of the dataset makes it more efficient to train the token frequency model while maintaining a similar shape of the frequency distribution.

To investigate the effect of token frequency, we augmented the training set using the normalised token frequencies. We achieved this by duplicating each verb in the dataset until its occurrence matched the indicated normalised token frequency. This augmentation of the dataset with normalised token frequency was done after dividing the data into a training and development set, and only the training set was augmented with token frequency. The development set was not augmented with the token frequency, because we evaluate real verb performance based on accuracy on verb types, which will be further explained in

Section 4.4.1. Before training any model, the training set was always randomly shuffled again.

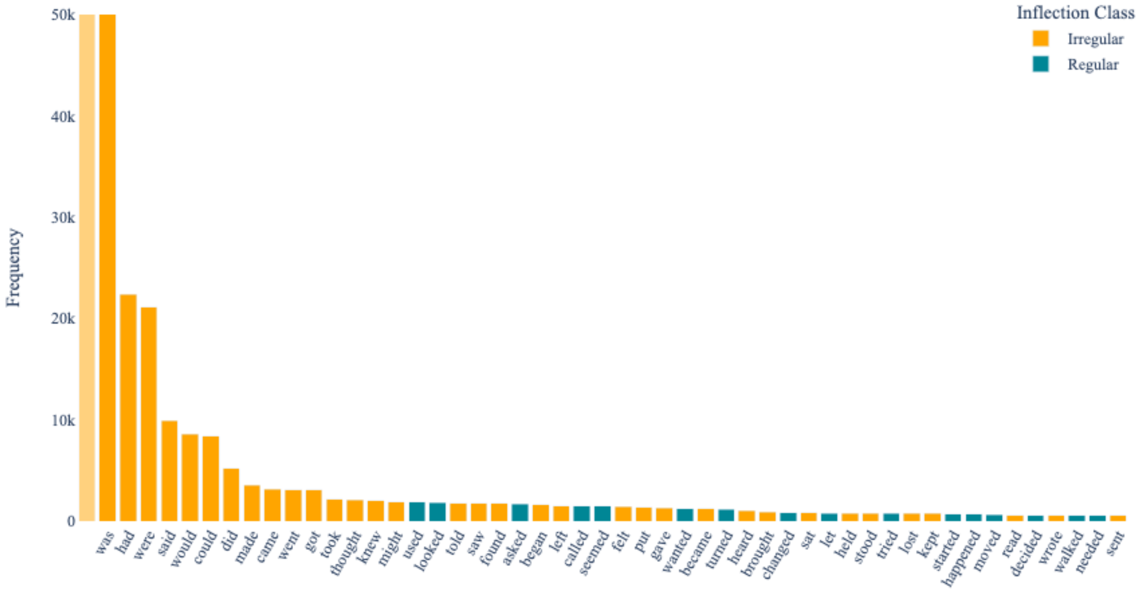


Figure 4.1: Top 50 most frequent verbs in the final dataset ordered on the raw token frequency of the past tense form. Note that the first verb *was* is spread out over two bars.

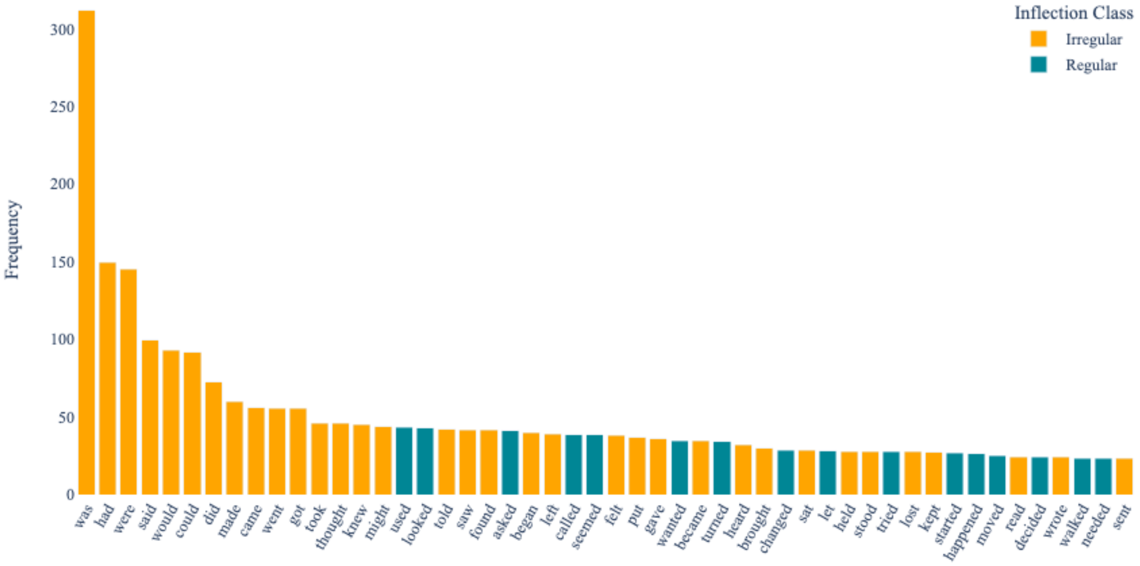


Figure 4.2: Top 50 most frequent verbs in the final dataset ordered on the square-root normalised token frequency of the past tense form.

4.1.5 Data Division

To train and evaluate the models based on the verb data, 80% of the real verbs were randomly selected as the training set, while the remaining 20% was allocated to the development set. As explained in Section 2.3.1, we follow the approach of Corkery et al. (2019) in considering the Wug Test the test of generalisation. Consequently, we divided the entire dataset of real verbs into a training set and a development set, without creating a separate test set. As mentioned in Section 3.4.4, the main reason for splitting the real verbs into training and development sets is to allow the use of an early stopping criterion that tracks the loss on the development set, helping to mitigate the potential problem of overfitting.

4.1.6 Data Statistics

The described data modifications result in a dataset that contains 4117 verb types in total. In this set, ∼5% of the types are labelled as irregular. As mentioned above, this set of verb types is divided into 80% training set (3292) and 20% development set (824). As explained, to take token frequency into account, the training set is augmented with the normalised token frequency distribution. The total sum of all verb token frequencies in the training set is 12.895, of which ∼20% is labelled as irregular. The difference in the size of the training set is given in Figure 4.3 and Table 4.2.

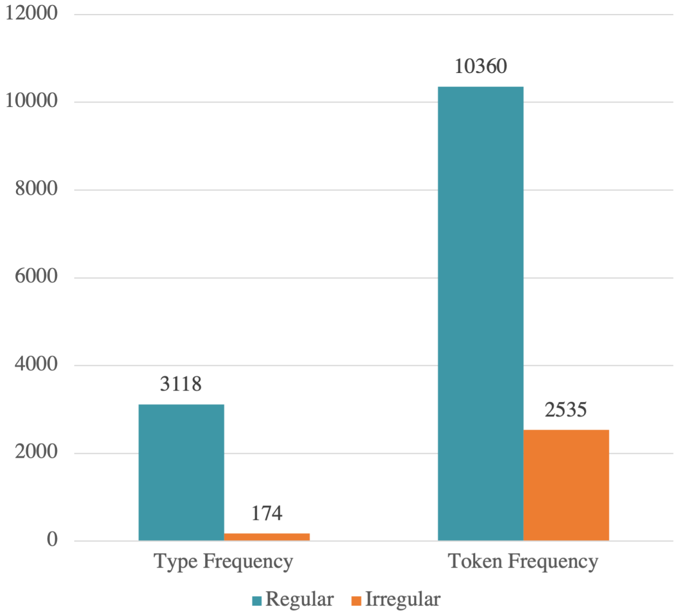


Figure 4.3: Size of the regular and irregular class in the training set based on type frequency and token frequency.

	Total	Irregular	Regular
Type frequency	3292	3118 (95%)	174 (5%)
Token frequency	12.895	10.360 (80%)	2535 (20%)

Table 4.2: Size of the regular and irregular class in the training set containing type frequency and token frequency.

4.2 Models

4.2.1 Model Description

To find answers to the first two research questions posed in Section 1.1, we compared four different models that differ in the following two aspects:

1. Type frequency vs Token frequency models:
    - Type frequency models are trained on a dataset in which each verb occurs once, representing type frequency (like previous studies).
    - Token frequency models are trained on a dataset that is augmented with the token frequency distribution (as described above in Section 4.1.4).
2. Single Task vs. Multi-task models:
    - The single task trained models only predict past tense forms.
    - The multi-task models have an auxiliary task in addition to the main task of predicting the past tense form: predicting whether verbs are regular or irregular. The implementation is further explained in the next section.

This results in four models:

1. $\texttt{TYPE}_{\texttt{SGT}}$: Type frequency model, single task training
2. $\texttt{TOKEN}_{\texttt{SGT}}$: Token frequency model, single task training
3. $\texttt{TYPE}_{\texttt{MTSK}}$: Type frequency model, multi-task training
4. $\texttt{TOKEN}_{\texttt{MTSK}}$: Token frequency model, multi-task training

As explained in the Theoretical Background (2) and Related Literature (3), the rationale for using token frequency is that it represents the linguistic input of humans more, and it may carry additional information to learn the productivity of the past tense inflection rules. The rationale for the multi-task training setup is to compel the model towards distinguishing between regular and irregular verbs. Since the dual-route view of P&P argues that regular and irregular verbs are handled separately, this comparison provides insight into the plausibility of this theory.

Two experiments were conducted with these four models to answer the third and fourth research questions. Therefore, this section and the next section first give a global overview of the implementation, training, and evaluation of these models.

4.2.2 Model Implementation and Architecture

To implement the models, Keras Tensor Flow (version 2.10.0) was used (Abadi et al., 2015). The four models have the same overall architecture, for which we refer to Section 2.4. At the end of Experiment 1, final model configuration choices are presented for each model type individually (Section 5.1.7).

Since we added an auxiliary task in the multi-task training setup, an additional dense layer predicts the verb class based on the context vector (the output of the attention mechanism) and decoder output. We give the formal definition of this binary classification layer below:

$$P(y|\mathbf{h}^{\text{dec-2}}, \mathbf{c}) = sigmoid(W \cdot [\mathbf{h}^{\text{dec-2}}; \mathbf{c}] + b)$$

Where $P(y|\mathbf{h}^{\text{dec-2}}, \mathbf{c})$ is the probability of the label y given the decoder's final hidden state $\mathbf{h}^{\text{dec-2}}$ and the context vector $\mathbf{c}$ from the attention mechanism. W is the weight matrix of the dense layer, b is the bias term, and $[\mathbf{h}^{\text{dec-2}}; \mathbf{c}]$ represents the concatenated vector formed by the context vector and the decoder output at the final time step.

4.3 Model Training

To make training efficient, teacher forcing was used; the decoder predicts the next characters based on the correct previous character instead of the previously predicted character (Sutskever et al., 2014). We used the Adam optimizer (Kingma & Ba, 2014). The learning rate, dropout rate, and batch size are selected in Section 5.1.7, based on the results of Experiment 1.

An early stopping criterion was used to prevent long training times and overfitting. The patience parameter of this early stopping criterion is selected and presented in Section 5.1.7 as well. As mentioned in the Related Literature (3), the previous studies Corkery et al. (2019) and K&C trained their models for a fixed number of epochs that led to near-perfect training accuracies. As will be further explained, the results of Experiment 1 revealed that the model tends to overfit on real verbs from the training data when too many training epochs were used as a result of a higher patience value. This problem of overfitting was mitigated by choosing the appropriate settings for the early stopping criterion.

In the single task training setting, the sparse categorical cross-entropy loss was computed based on the predicted sequence of characters. The formal definition of this is given below, where N is the number of predicted characters in the sequence, p_i is the predicted chance of the correct character in position i, and y_i is the true character in that position:

$$L_{\text{seq}} = -\frac{1}{N}\sum_{i=1}^{N} \log(p_i[y_i])$$

For the auxiliary task in the multi-task training setup, the binary cross-entropy loss was computed. The formal definition of this is given below, where y_{class} is the correct class of the sequence and p_{class} is the prediction chance of the correct class.

$$L_{\text{class}} = -\left(y_{\text{class}} \cdot \log(p_{\text{class}}) + (1 - y_{\text{class}}) \cdot \log(1 - p_{\text{class}})\right)$$

In this multi-task training setup, the overall loss is a weighted sum of the loss on the character predictions and the label class prediction. Since the character prediction is the main task, the weight of the loss on this task is larger than that of the auxiliary task. In Section 5.1, we show our investigation of three options for this weighted sum. Below this weighted sum is given, where W_1 is the weight for the main task (character prediction), and W_2 is the weight for the auxiliary task (class prediction):

$$L_{\text{total}} = W_1 \cdot L_{\text{seq}} + W_2 \cdot L_{\text{class}}$$

4.4 Model Evaluation

4.4.1 Real Verb Accuracy

We evaluated the models' performance on real verbs by computing the accuracy of inflection predictions. This was done by taking the number of completely correctly predicted verb inflections divided by the total number of verbs of the set. For both the type and token frequency models, the accuracy is always based on the number of verb *types*. This is because, for evaluation, we consider each verb equally important, whereas basing accuracy on token frequency would give more weight to frequently occurring verbs and less to those that appear less often. We also computed the accuracy for the regular and irregular class individually. In the exact same way, we computed the accuracy of the class label predictions of the multi-task training models.

4.4.2 Nonce Verb Correlation

As explained in Section 2.3.1, we utilised Wug Test nonce verbs as test set, which allows to evaluate to what extent model generalisations are human-like. To evaluate model performance on these nonce verbs, we followed Corkery et al. (2019) and K&C by taking the correlation between the model predictions and those of A&H's human Wug experiments.

For the prediction of nonce verb inflections, we applied a beam search algorithm that predicts the top 12 outputs of the model. Subsequently, each prediction from the beam was categorised into one of the four nonce verb inflection classes as indicated by A&H: *Regular, Irregular 1, Irregular 2*[2] or *Other*. We assigned these categories the associated beam probability. We normalised these probabilities such that the total sum of probabilities for each nonce verb is always 100%. The result for one nonce verb may look like the example in Table 4.3.

	Regular	Irregular 1	Irregular 2	Other
	spliŋ	spluŋ	splæŋ	...
Beam	2%	94%	3%	1%

Table 4.3: Example of beam probability predictions for one run and one nonce verb *spling*.

Finally, we computed the correlation between these model probabilities and the human production probabilities from the A&H experiment. In line with previous studies (A&H; Corkery et al., 2019; K&C) and for completeness, we always computed both the Spearman (monotonic) and Pearson (linear) correlation coefficients.

[2] A&H only provided an *Irregular 2* form for 11 of the 58 nonce verbs.

4.4.3 Aggregate Results

Corkery et al. (2019) observed variation in results when running the exact same model multiple times. Therefore, they proposed aggregating the predictions from multiple runs of the same model. In this approach, each individual model run is seen as an individual speaker performing the Wug Test. We followed Corkery et al. (2019) in this. We aggregated the beam probability results over multiple runs per model[3]. This means that the sum of all beam probabilities per class per nonce verb was taken and then divided by the number of runs of the model again. As indicated in Experiments 1 and 2, we present aggregate results based on five runs per model. In our Chapter of Final Results (6), we present aggregate results from the two experiments.

[3] Where each *run* means that all settings are the same, except for random weight initialization and random shuffle of training data.

Chapter 5

Experiments

5.1 Experiment 1: Model Selection

5.1.1 Objective and Procedure

To find an answer to our third research question, we investigated the effect of different model configuration choices on the performance of the four models on real and nonce verbs. So far, previous studies have only worked with models similar to our $\texttt{TYPE}_{\texttt{SGT}}$ model. As mentioned, these studies do not report any exploration of data and model settings for this exact task. Moreover, since our four models differ from each other in training data and/or task setting, each model may require different training and architecture settings for optimal performance.

We used a set of *default settings*, given in Table 5.1 unless indicated otherwise in the steps of this experiment. Like Corkery et al. (2019), we adopted the LSTM layer number and size, embedding layer size, and dropout rate from K&C. We set the default batch size to 32. The default learning rate is 0.001, which is the default learning rate of the Adam optimizer in Keras Tensor Flow (Abadi et al., 2015; Kingma & Ba, 2014).

Hyperparameter	Setting
Number of LSTM layers	2
LSTM size	100
Embedding size	300
Dropout rate	0.3
Batch size	32
Learning rate	0.001
Patience	15

Table 5.1: Default settings for both the encoder and decoder components of the models.

In this experiment we investigated the effect of the following five categories sequentially:

- 5.1.2 Data modifications
- 5.1.3 Number of Layers
- 5.1.4 Weighted sum for the overall loss in the multi-task training setup

- 5.1.5 Early stopping criterion
- 5.1.6 Hyperparameter settings: layer sizes, dropout, learning rate, batch size

In this experiment, we trained each model five times, after which average accuracies and aggregate Wug Test correlations were computed for evaluation (as explained in more detail in Section 4.4). By iteratively selecting the best model after each step, we finally obtained a selection of the best model for each of the four models (Section 5.1.7). Because we consider the correlation with human results the indicator of how well a model generalises and fits with human behaviour on novel input, we consistently relied on these scores for our model selection.

5.1.2 Data Modifications

As described in Section 4.1.2.2, it is unknown what the effect is of including the set of verbs that were omitted from A&H's verb data, which is the dataset that was used by Corkery et al. (2019) and K&C as well. We also do not know the effect of excluding *doubles* for the verbs that have more than one correct inflection. Therefore, the first step of this experiment is to explore model performance on real and nonce verbs when including and excluding these verbs.

In Table 5.2, the average accuracy on real verbs and the correlation between the aggregate model predictions and human Wug Test data are presented for each of the four models. From these results, we can conclude that all four models have the best fit with human data when both the missing verbs and *doubles* are included in the data. However, this did not yield optimal training and development accuracies. In fact, it led to the lowest accuracy on irregular verbs from the training set for both type frequency models.

Excluding *doubles* by only keeping the form with the highest token frequency generally results in lower accuracies on the irregular verbs from the development set. However, when the *doubles* are included, the chances of predicting the correct verb inflection are bigger, as there is more than one correct inflection for the *doubles*. This flexibility in correctness explains the difference in accuracy, as we found that the accuracy on all other verbs that have only one inflection did not drastically differ.

Both the missing verbs and the *doubles* were included in the final dataset because their inclusion yielded the highest correlation with the A&H human data. This is also in line with the fact that the inclusion of both also makes the input data most complete, compared to the input of human language learners. Though this makes sense, it should be noted that: (1) The set of missing verbs contains highly frequent highly irregular verbs (see Table 4.1, Section 5.1.2). Even though this can pose a challenge for the model's performance on real verbs, assuming that these forms are rather memorised than generalised by humans, their inclusion led to a stronger fit with human speaker inflections of nonce forms. (2) The linguistic context of humans allows for lexical disambiguation between *doubles* such as *ring* and *wring*, which could be considered an advantage that our models do not have. Nevertheless, representing the phenomenon of occasionally having more than one correct past tense form for the same present tense form still led to a stronger fit with human spekaer inflections of nonce forms.

			Training			Development			
Missing	Doubles	Overall	Irregular	Regular	Overall	Irregular	Regular	r	ρ
TYPE$_{SGT}$									
incl.	**incl.**	**98.75**	**81.12**	**99.75**	**95.56**	**43.18**	**98.29**	**.56**	**.51**
incl.	excl.	99.41	92.66	99.76	95.55	30.86	98.23	.54	.49
excl.	incl.	99.30	89.82	99.80	95.99	43.68	98.49	.48	.50
excl.	excl.	99.14	85.03	99.86	96.00	30.63	98.71	.52	.52
TOKEN$_{SGT}$									
incl.	**incl.**	**99.31**	**93.60**	**99.64**	**95.44**	**48.18**	**97.94**	**.55**	**.48**
incl.	excl.	99.66	96.57	99.83	95.26	28.00	98.08	.52	.48
excl.	incl.	99.06	92.72	99.41	95.01	43.16	97.58	.49	.47
excl.	excl.	99.72	98.36	99.79	95.78	37.50	98.18	.48	.42
TYPE$_{MTSK}$									
incl.	**incl.**	**96.76**	**45.96**	**99.62**	**95.41**	**37.27**	**98.42**	**.62**	**.60**
incl.	excl.	98.00	67.22	99.67	95.56	24.57	98.51	.57	.55
excl.	incl.	98.77	82.91	99.60	94.91	37.89	97.60	.49	.48
excl.	excl.	98.54	74.34	99.79	95.73	31.88	98.33	.53	.55
TOKEN$_{MTSK}$									
incl.	**incl.**	**99.26**	**94.16**	**99.55**	**94.42**	**45.00**	**96.98**	**.47**	**.42**
incl.	excl.	99.07	92.07	99.46	94.91	32.00	97.60	.44	.41
excl.	incl.	99.16	92.00	99.53	94.77	37.37	97.43	.31	.39
excl.	excl.	99.46	93.21	99.78	95.46	37.50	97.80	.28	.40

Table 5.2: Average accuracy on real verb data and Pearson's r and Spearman's ρ with human Wug Test data: with and without missing verbs and *doubles*.

5.1.3 Number of Layers

In addition to our architecture with two stacked LSTM layers in both the encoder and decoder, we explored a simpler architecture with a single LSTM layer in each. Using a single LSTM layer instead of multiple layers reduces the model's computational complexity and can help mitigate overfitting risks. As can be seen from Table 5.3, the average accuracy on the training and development sets are relatively consistent for both settings. We only note small differences in accuracy on the irregular class for all models, except TOKEN$_{SGT}$. The correlations between the model and human nonce verb inflections are higher or similar for each model when using two LSTM layers instead of one. Therefore, we maintained the architecture with two LSTM layers for each of the four models.

5.1.4 Multi-task Training Weighted Sum Overall Loss

In this third step, we investigated different settings of the weighted sum of the loss for the multi-task training models. In addition to a weighted sum of (0.7, 0.3) for the main and auxiliary loss, respectively, we explored two other options: (0.8, 0.2) and (0.6, 0.4). In all three options, the model's loss on the main task has the largest weight since this task should remain the main focus of the model during training.

The different weightings shift the model's focus between two tasks: accurately inflecting verbs and correctly classifying them as regular or irregular. Assigning a higher weight to the auxiliary classification task puts more emphasis on the distinction between regular and irregular verbs. If this auxiliary task carries potential benefits for the main task of verb

	Training			Development				
Num layers	Overall	Irregular	Regular	Overall	Irregular	Regular	r	ρ
TYPE$_{\text{SGT}}$								
1	98.52	76.18	99.80	95.78	45.46	98.39	.53	.48
2	**98.75**	**81.12**	**99.75**	**95.56**	**43.18**	**98.29**	**.56**	**.51**
TOKEN$_{\text{SGT}}$								
1	99.22	93.48	99.56	95.07	48.64	97.66	.46	.42
2	**99.31**	**93.60**	**99.64**	**95.44**	**48.18**	**97.94**	**.55**	**.48**
TYPE$_{\text{MTSK}}$								
1	96.89	51.46	99.43	95.15	41.82	97.99	.60	.53
2	**96.76**	**45.96**	**99.62**	**95.41**	**37.27**	**98.42**	**.62**	**.60**
TOKEN$_{\text{MTSK}}$								
1	98.20	81.91	99.13	93.69	47.72	96.18	.47	.42
2	**99.26**	**94.16**	**99.55**	**94.42**	**45.00**	**96.98**	**.47**	**.42**

Table 5.3: Average accuracy on real verb data and Pearson's r and Spearman's ρ with human Wug Test data: using different numbers of LSTM layers.

inflection, but too little weight is assigned to the auxiliary task loss, the model may not focus enough on the auxiliary task to benefit from auxiliary task. Conversely, assigning the auxiliary loss a too-heavy weight could generally detract from the model's ability to perform the primary task of verb inflection, leading to poorer inflections. If the auxiliary task is not beneficial for the model's performance on the main task at all, it is also expected that more weight to the auxiliary task loss leads to a decrease in main task performance. Therefore, we aim to strike the right balance between the weights of the main and auxiliary task loss.

As can be seen from Table 5.4, both models obtained the highest correlation with human Wug Test data with a weighted sum of (0.7, 0.3) for the main and auxiliary task loss, respectively. For `TOKEN`$_{\text{MTSK}}$, this is also the model that obtained the highest main task accuracy on the training set. This is not the case for `TYPE`$_{\text{MTSK}}$, which obtained a lower accuracy on the training verbs, which is due to poorer performance on the irregular class. For both models, the main task accuracy on the development set is relatively similar across the three settings.

Furthermore, we see for both models that the weakest performance on the auxiliary task is obtained when the Wug Test correlations are the strongest. This suggests an inverse relationship: models that better align with human nonce verb inflections tend to show lower accuracy in verb class prediction. This could imply that the model achieves better generalisation when it prioritises less its performance on the auxiliary task. However, if this were the case, we would expect that the model would have obtained the best generalisations and the weakest auxiliary task accuracies when the auxiliary task is given the lowest weight (i.e., 0.2). However, this is not the case.

In Chapters 6 and 7, we discuss in more detail the effect of the auxiliary task by comparing their overall performance to those of the models without auxiliary task and the implications of these results. For our model selection, we still prioritised performance on the main task, since it is the objective to select models with nonce verb generalisations that are most similar to those of humans. Therefore, we maintained the weighted sum of (0.7, 0.3) for the computation of the models' overall loss during training.

Weighted sum	Training			Development					Auxiliary	
Main/Aux	Overall	Irregular	Regular	Overall	Irregular	Regular	r	ρ	Train	Dev
TYPE$_{\text{MTSK}}$										
0.6/0.4	97.12	52.36	99.64	95.59	37.73	98.56	.41	.46	82.43	81.77
0.7/0.3	**96.76**	**45.96**	**99.62**	**95.41**	**37.27**	**98.42**	**.62**	**.60**	**27.23**	**27.08**
0.8/0.2	98.21	74.04	99.57	95.07	43.64	97.76	.38	.43	89.31	88.20
TOKEN$_{\text{MTSK}}$										
0.6/0.4	97.75	85.28	98.39	93.20	46.36	95.63	.29	.37	94.05	91.36
0.7/0.3	**99.26**	**94.16**	**99.55**	**94.42**	**45.00**	**96.98**	**.47**	**.42**	**72.30**	**70.66**
0.8/0.2	99.06	91.34	99.50	95.15	47.73	97.61	.40	.40	85.83	84.90

Table 5.4: Average accuracy on real verb data and Pearson's r and Spearman's ρ with human Wug Test data, as well as average accuracy on the auxiliary task class predictions.

5.1.5 Early Stopping Criterion

For the fourth step of this experiment, we explored using an early stopping criterion. An early stopping criterion monitors the overall loss during training and halts training if no improvement is observed after a predefined number of epochs, referred to as the patience threshold. The previous studies of Corkery et al. (2019) and K&C used a fixed number of training epochs (100) to train their models, leading to almost perfect training accuracy. Though the expectation of a model performing well on training verbs is sensible—given that humans also perform well on their acquired verbs—overfitting may have been a problem in the previous studies. As mentioned in the Related Literature Chapter (3), further investigation by Corkery et al. (2019) showed that optimal correlations with human nonce verb inflections were obtained after only 10 training epochs instead of 100. An early stopping criterion can prevent training on too many epochs, making training times more efficient and potentially improving generalisation to nonce verbs by preventing the potential overfitting.

To explore this option, we not only considered using different patience values but also investigated the difference between tracking loss on the *training* set and the *development* set. Although tracking training loss is not a commonly used method, we consider striving for optimal performance on the training set reasonable, because cognitive models may be expected to mimic human behaviour on real verbs as well, which would be a nearly-perfect performance. Unlike previous studies, we ensure not to train longer than necessary to achieve the desired results by implementing an early stopping criterion instead of relying on a fixed number of epochs.

Another reason to explore tracking training loss with an early stopping criterion is that this allows the use of all real verbs as training data (cf. Corkery et al., 2019). Not leaving out 20% of the data is more representative of the task of human language learning. However, we argue that we must also consider tracking loss on a development set, regardless of the downside of some real verbs being omitted from the training data. This is for the conventional reason that the loss on a development set indicates when training no longer improves the generalisation of examples that are not seen during training, which is the actual goal and cannot be tracked based on training loss.

In Table 5.5, we present for each of the four models the results of tracking training and development loss with an early stopping criterion using different patience threshold values. The epoch number shown indicates the epoch to which network weights were restored due to no further improvement within the specified patience period. The results show that a

	Training			Development					Num
Patience	Overall	Irregular	Regular	Overall	Irregular	Regular	r	ρ	Epochs
TYPE$_{SGT}$									
10 Dev	98.10	68.88	99.74	95.88	40.91	98.72	.47	.48	49
15 Dev	**98.75**	**81.12**	**99.75**	**95.56**	**43.18**	**98.29**	**.56**	**.51**	**54**
20 Dev	99.42	91.91	99.86	96.10	47.72	98.62	.48	.45	68
25 Dev	99.60	94.27	99.89	95.90	48.18	98.42	.49	.50	67
10 Train	99.91	99.44	99.93	95.95	49.09	98.34	.46	.48	103
15 Train	99.95	100.0	99.94	95.75	44.55	98.44	.40	.42	109
20 Train	100.0	100.0	100.0	96.07	51.18	98.39	.41	.32	131
25 train	100.0	100.0	100.0	95.94	49.18	98.42	.39	.37	156
TOKEN$_{SGT}$									
10 Dev	99.41	96.07	99.59	94.66	46.82	97.19	.50	.45	18
15 Dev	99.31	93.60	99.64	95.44	48.18	97.94	.55	.48	17
20 Dev	**99.36**	**93.82**	**99.68**	**95.21**	**48.18**	**97.74**	**.61**	**.53**	**22**
25 Dev	99.53	95.51	99.75	95.32	47.27	97.89	.53	.50	30
10 Train	99.93	99.21	99.97	96.17	51.36	98.57	.50	.47	48
15 Train	99.98	99.89	99.99	95.85	45.91	98.49	.51	.48	83
20 Train	100.0	100.0	100.0	95.87	49.43	98.34	.55	.44	96
25 train	99.99	100.0	99.99	95.72	47.27	98.24	.52	.42	97
TYPE$_{MTSK}$									
10 Dev	96.96	52.81	99.44	95.22	37.27	98.22	.48	.48	32
15 Dev	**96.76**	**45.96**	**99.62**	**95.41**	**37.27**	**98.42**	**.62**	**.60**	**40**
20 Dev	97.96	68.54	99.61	95.66	43.64	98.37	.52	.50	49
25 Dev	98.06	74.04	99.40	95.07	43.64	97.79	.42	.45	77
10 Train	95.66	92.25	99.87	95.66	45.91	98.29	.41	.47	80
15 Train	99.94	99.10	99.99	95.73	46.36	98.31	.42	.41	102
20 Train	99.99	99.89	100.0	95.80	44.55	98.49	.40	.37	147
25 Train	99.99	99.89	100.0	95.51	45.45	98.14	.48	.39	196
TOKEN$_{MTSK}$									
10 Dev	99.21	92.47	99.60	95.05	44.54	97.81	.46	.47	23
15 Dev	99.26	94.16	99.55	94.42	45.00	96.98	.47	.42	19
20 Dev	**99.12**	**93.25**	**99.43**	**94.64**	**45.91**	**97.24**	**.54**	**.47**	**23**
25 Dev	99.10	94.27	99.38	94.54	47.27	97.09	.53	.44	22
10 Train	99.98	100.0	99.98	95.24	46.82	97.76	.45	.43	61
15 Train	99.99	99.89	99.99	95.73	46.81	98.27	.46	.40	115
20 Train	99.99	100.0	99.99	95.85	51.36	98.22	.46	.39	113
25 Train	99.99	99.89	99.99	95.92	46.82	98.54	.46	.45	140

Table 5.5: Average accuracy on real verb data and Pearson's r and Spearman's ρ with human Wug Test data and average number of epochs: early stopping criterion tracking training or development set loss with different patience parameters.

patience value of 20 while tracking development loss led to the highest correlations with human Wug Test data for TOKEN$_{SGT}$ and TOKEN$_{MTSK}$, and a patience of 15 while tracking development loss for TYPE$_{SGT}$ and TYPE$_{MTSK}$. For this reason, we selected the models' early stopping criterion settings leading to optimal correlations.

As evident from the results, tracking the training loss results in a nearly perfect accuracy on the training data for all four models, which aligns with findings from previous studies. However, when the development loss is tracked instead of the training loss, the model's accuracy on the training data notably drops across all cases, particularly for both type frequency models on the irregular class. This is because minimal loss on the development set is reached much earlier on the training set. Consequently, the models train for a smaller number of epochs when development loss is tracked. On the other hand, tracking development loss generally does not appear to affect overall accuracies on the development set. Most importantly, it even results in notably stronger correlations with the human Wug Test nonce verb inflections. These results suggest that tracking development loss mitigates the problem of overfitting on the training set.

For the type frequency models, a significantly smaller number of training epochs—hence generally less exposure to irregular forms—resulted in lower performance on the irregular forms of the training set. Token frequency models, on the other hand, have proportionally more exposure to irregular forms during training. This could explain their more robust performance on irregular training verbs, regardless of shorter training times. This suggests that for type frequency models, there may be a trade-off between capturing human behaviour on real verbs and capturing human behaviour on nonce verbs, whereas this is not the case for token frequency models. We discuss this finding further in the Results and Discussion chapters (6, 7).

5.1.6 Hyperparameter Settings

In this final part of the first experiment, we explored different combinations of hyperparameters. In the first step, we look at the effect of the size of the network in combination with the dropout rate. In addition to the described default size settings of 300 units per the embedding layer and 100 units per the LSTM layer, we explored layer size combinations of 400 and 200 units for the embedding and LSTM layers, respectively. These two options are combined with dropout rates of 0.1 and 0.5, in addition to 0.3. By exploring options for the dropout rate and layer sizes at the same time, we attempt to capture the nuanced interplay between regularization strength and the model's representational capacity.

In the next step, we explored two other batch sizes in addition to the default of 32: 16 and 64. The models differ from each other in training set size and frequency distributions; therefore, they might benefit differently from different batch sizes. We did this by selecting the two best models so far and testing the two alternative batch sizes with the selected models. Finally, we investigated the performance of models with a larger (0.01) and smaller (0.001) learning rate compared to the models with the default learning rate (0.0001). Again, we did this by testing these alternative settings on the two best models so far.

In Appendix A, all results of this final step of Experiment 1 are presented. In Tables 5.6 and 5.7, we show the final selection of models and their results. As can be seen from these tables, the optimally performing models on the Wug Test correlations differ from each other in hyperparameter settings. As observed before in this experiment, we saw a further drop in the performance of the type frequency models on the irregular verbs from the training set as a result of using settings that led to the strongest Wug Test correlations.

5.1.7 Model Selection Conclusion

In this first experiment we investigated the effect of different data, training and architecture configurations for each of the four models, $\texttt{TYPE}_{\texttt{SGT}}$, $\texttt{TOKEN}_{\texttt{SGT}}$, $\texttt{TYPE}_{\texttt{MTSK}}$, $\texttt{TOKEN}_{\texttt{MTSK}}$ and

	Patience	Embedding/LSTM size	Dropout	Batch size	Learning rate
TYPE$_{SGT}$	15	400/200	0.5	64	0.01
TOKEN$_{SGT}$	20	300/100	0.3	32	0.001
TYPE$_{MTSK}$	15	300/100	0.3	32	0.01
TOKEN$_{MTSK}$	20	300/100	0.5	16	0.001

Table 5.6: Settings of the finally selected models for each model type.

	Training			Development				
	Overall	Irregular	Regular	Overall	Irregular	Regular	r	ρ
TYPE$_{SGT}$	97.05	57.53	99.24	95.44	41.36	98.24	.61	.59
TOKEN$_{SGT}$	99.36	93.82	99.68	95.21	48.18	97.74	.61	.53
TYPE$_{MTSK}$	97.13	53.26	99.58	95.32	36.82	98.27	.66	.61
TOKEN$_{MTSK}$	99.00	92.25	99.39	95.39	50.00	97.86	.58	.48

Table 5.7: Results of the finally selected models for each model type. Average accuracies on real verbs and aggregate correlations on nonce verbs (Pearson's r and Spearman's ρ).

selected for each model the version with the highest correlation with the A&H human data on the Wug Test.

All four models have in common that the highest correlations were obtained when we included all verbs in the dataset: both the *doubles* and *missing verbs*. As mentioned, this is also the most human-like input. It is also true for all four models that two layers of LSTM yielded the strongest correlations compared to only one LSTM layer. The results of this experiment also showed that tracking the development loss with an early stopping criterion is effective in helping prevent overfitting on the training set and yielded the highest correlations with human nonce verb inflections. This is why we split the real verb data into a training and development set. For both multi-task models, a weighted sum of (0.7,0.3) worked best for the main and auxiliary tasks, respectively. Table 5.6 and 5.7 summarise the other settings that were finally selected and the obtained results.

In conclusion, the results of this experiment indicate that the type frequency models obtained the highest correlations with human Wug Test data, especially TYPE$_{MTSK}$ with correlations of .66 and .61 (Pearson and Spearman). TYPE$_{SGT}$, TOKEN$_{SGT}$, and TOKEN$_{MTSK}$ obtained similar Pearson correlations of around .60, but differ in strength of the Spearman correlation, with TOKEN$_{MTSK}$ performing worst. However, as mentioned throughout this experiment, the type frequency models with the highest Wug Test correlations obtained significantly worse accuracies on irregular verbs from the training set. A further comparison between the four models and discussion of our observations in this Experiment are given in the Results and Discussion chapters (6, 7).

5.2 Experiment 2: Impact of Verb Distribution

5.2.1 Description of the Experiment

In this second experiment, we explored the effect of verb distribution across the training and development sets. As previously explained, 20% of the real verbs served as development sets, so overfitting can be prevented by using an early stopping criterion that tracks development loss. While the results from the first experiment (5.1.5) indicate that this method effectively improves the models' alignment with human behaviour on nonce verbs, it can be questioned whether the distribution of real verbs between the training and development sets influences model performance.

Though it is a common approach in the field of machine learning to randomly select one development set, it is for the task at hand not as straightforward. In order to appropriately measure the ability of the model to generalise to novel forms, the development set must be representative of the training data. However, in English past tense inflection, it is complex to determine which set of forms suits best as development set to form a representative set of generalisation. When it comes to highly or completely idiosyncratic past tense forms, it is nearly impossible to predict the correct inflection without prior exposure to the correct inflection, because there is no pattern in the training data from which it could generalise to these forms. Additionally, for the irregular verbs that can also be considered semi-regular, it is a challenging task to determine which set is a representative development set containing examples of verbs from cohorts of similarly inflected verbs (e.g., {*sing—sang, ring—rang*, ...}).

Moreover, it is also relevant that the training set contains sufficient information to generalise appropriately to our test set, the Wug Test nonce verbs. One random distribution of training-development data may be more suitable for human-like generalisations to the nonce verbs than another. It may also vary across the four models what works best, given the differences in task and data. For the token frequency models, different splits could lead to more pronounced changes in the models' learning outcomes. This is because there is a set of irregular verbs that occur very frequently, hence different splits lead to bigger differences in the verb frequency distribution than when all verbs occur equally frequently in the training set. At the same time, token frequency models always have proportionally more exposure to the irregular forms during training compared to type frequency models, which could also make them more resilient to the omission of a few verb types.

To investigate whether the distribution of verbs over the training and development sets has an effect on the performance on real verbs and nonce verbs, we trained and tested our selected models from the previous experiment on four alternative training-development splits. We distributed the real verb data such that each verb type is once part of the development set and four times part of the training set, as visualised in Figure 5.1. Like before, the training data of the token frequency models is augmented with the token frequency distribution after splitting the data into a training and development set. On each of the four additional splits, we executed the same procedure as on the split used in Experiment 1: we ran all models five times, meaning that each model is trained and evaluated five times, after which the average accuracy on real verbs and aggregated correlation with human nonce verb inflections were computed.

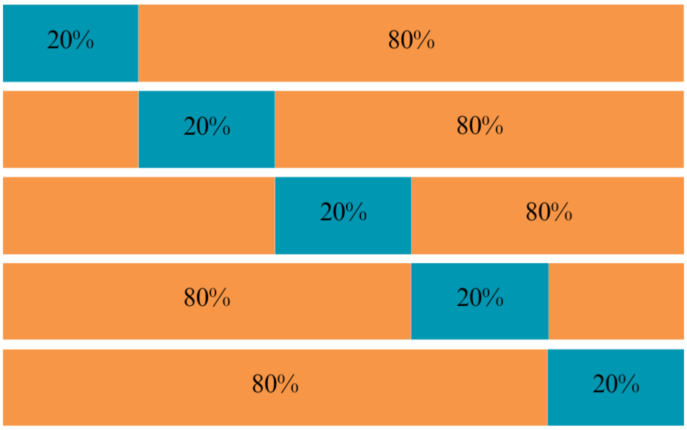


Figure 5.1: Five splits of real verb data split into a training set (orange) and a development (blue) set.

5.2.2 Influence of Distribution on Real Verb Accuracy

In Table 5.8, we show the average accuracy on the training and development verbs of the four models on each split. The results on Split 1 are the results from Experiment 1. In Table 5.9, the standard deviation of the average accuracy on each split is shown for the four models. This gives insight into how much each model varied in performance *between* the five splits. We focus in this section on model performance per split; for further overall comparison between the four models we refer to Chapter 6.

The first thing that can be noted from Table 5.9 is that the performance of TOKEN$_{\text{SGT}}$ is most stable throughout the five different splits on all categories. Secondly, TOKEN$_{\text{MTSK}}$ has a higher overall accuracy standard deviation on the training and development set than the other models, followed by TYPE$_{\text{MTSK}}$. TOKEN$_{\text{MTSK}}$ is also the only model that shows much higher variation between splits on the regular class accuracy on both the training and development set. These higher standard deviations of TOKEN$_{\text{MTSK}}$ seem to mainly come from its poor performance on *one* of the splits: Split 3 (Table 5.8).

Furthermore, the standard deviations presented in Table 5.9 demonstrate that the performance of the type frequency models on the irregular training verbs is less stable across splits compared to the token frequency models. However, the type frequency models have high standard deviations on the training irregular accuracy *within* most splits as well, as can be seen in Table 5.8. Hence, they seem to obtain less stable accuracies on this class in general. Nonetheless, the standard deviation on irregular training verb accuracy massively differs between splits for the type frequency models. For instance, on the irregular training verbs, TYPE$_{\text{SGT}}$ has a standard deviation of only 0.99 (with an accuracy of 95.15%) using Split 3, while the standard deviation is 12.87 (with an accuracy of 61.38%) using Split 4. This suggests that a different distribution of verbs can heavily influence the type frequency models' behaviour on the irregular training verbs. This is in line with the type frequency models' usually lower performance on the irregular training verbs. This weakness of the type frequency model can be explained by their proportionally lower exposure to irregular verbs during training compared to the token frequency models.

In conclusion, the overall effect of using a different distribution of verbs over the training and development set is weak for most models' overall accuracies on real verbs, except for TOKEN$_{\text{MTSK}}$, which mostly had exceptionally lower accuracies on one of the splits. The type

	Training			Development		
	Overall	Irregular	Regular	Overall	Irregular	Regular
TYPE_{SGT}						
Split 1	97.05% (0.51)	57.53% (9.82)	99.24% (0.27)	95.44% (0.43)	41.36% (4.37)	98.24% (0.44)
Split 2	97.64% (0.17)	67.54% (5.62)	99.39% (0.17)	94.37% (0.59)	33.33% (6.54)	97.27% (0.35)
Split 3	97.08% (0.60)	95.16% (0.99)	99.09% (0.39)	94.73% (0.58)	35.92% (4.91)	98.47% (0.69)
Split 4	97.12% (0.80)	61.38% (12.87)	99.06% (0.48)	94.95% (0.25)	38.75% (6.18)	98.17% (0.62)
Split 5	96.70% (0.71)	49.20% (11.41)	99.24% (0.24)	94.34% (0.20)	25.42% (4.75)	98.64% (0.33)
$\text{TOKEN}_{\text{SGT}}$						
Split 1	99.36% (0.49)	93.82% (5.21)	99.68% (0.24)	95.22% (1.04)	48.18% (4.93)	97.74% (0.83)
Split 2	99.74% (0.18)	98.80% (0.60)	99.80% (0.18)	94.44% (0.56)	49.23% (4.21)	96.62% (0.58)
Split 3	99.47% (0.48)	99.29% (0.62)	99.70% (0.30)	94.59% (0.72)	40.82% (5.20)	98.01% (0.51)
Split 4	99.33% (0.58)	95.52% (3.64)	99.55% (0.42)	94.42% (0.76)	47.92% (7.80)	97.25% (0.61)
Split 5	99.67% (0.23)	97.36% (3.11)	99.81% (0.13)	94.83% (0.35)	40.42% (3.78)	98.24% (0.55)
$\text{TYPE}_{\text{MTSK}}$						
Split 1	97.13% (0.36)	53.26% (4.77)	99.59% (0.16)	95.32% (0.32)	36.82% (5.18)	98.27% (0.24)
Split 2	98.31% (0.95)	73.01% (17.07)	99.74% (0.12)	94.90% (0.45)	38.46% (7.48)	97.54% (0.66)
Split 3	95.49% (1.30)	92.22% (1.56)	99.05% (1.09)	92.35% (3.21)	25.71% (5.12)	96.61% (3.32)
Split 4	98.56% (0.77)	78.62% (16.71)	99.67% (0.16)	94.81% (0.58)	41.25% (6.65)	97.97% (0.52)
Split 5	98.63% (1.09)	79.20% (16.98)	99.67% (0.26)	94.42% (0.48)	30.83% (4.52)	98.39% (0.61)
$\text{TOKEN}_{\text{MTSK}}$						
Split 1	99.00% (0.72)	92.25% (6.35)	99.39% (0.42)	95.39% (0.38)	50.00% (7.00)	97.86% (0.15)
Split 2	99.37% (0.53)	95.63% (2.81)	99.57% (0.40)	93.67% (0.67)	39.49% (5.62)	96.09% (0.61)
Split 3	90.46% (5.98)	88.60% (4.79)	92.38% (7.14)	86.99% (8.55)	27.35% (2.74)	90.80% (9.12)
Split 4	99.13% (0.60)	94.37% (5.68)	99.40% (0.34)	94.51% (0.40)	46.67% (6.69)	97.37% (0.75)
Split 5	98.10% (0.89)	88.16% (5.34)	98.66% (0.69)	93.30% (0.82)	37.50% (5.31)	96.83% (0.82)

Table 5.8: Average accuracy and standard deviation on training and development verbs, overall and per class of verbs (n=5 per split).

	Training			Development		
	Overall	Irregular	Regular	Overall	Irregular	Regular
TYPE_{SGT}	0.30%	15.67%	0.12%	0.41%	5.48%	0.47%
$\text{TOKEN}_{\text{SGT}}$	0.16%	2.04%	0.09%	0.30%	3.86%	0.58%
$\text{TYPE}_{\text{MTSK}}$	1.20%	12.68%	0.25%	1.04%	5.61%	0.64%
$\text{TOKEN}_{\text{MTSK}}$	3.40%	3.00%	2.77%	2.98%	7.88%	2.56%

Table 5.9: Standard deviation of the average accuracy per split.

frequency models also show on the irregular class of the training data that they do not only have much lower accuracies but also less stable results between and within different splits, compared to token frequency models. These results suggest that it would be beneficial for type frequency models to dive deeper into the investigation of which exact set of irregular verbs should be in the training and development set, in order to capture human behaviour on real verbs. However, we consider this task beyond the scope of this thesis and continued by aggregating the results from all five splits, so that we obtain the most representative result within our possibilities. We show these overall results in Chapter 6.

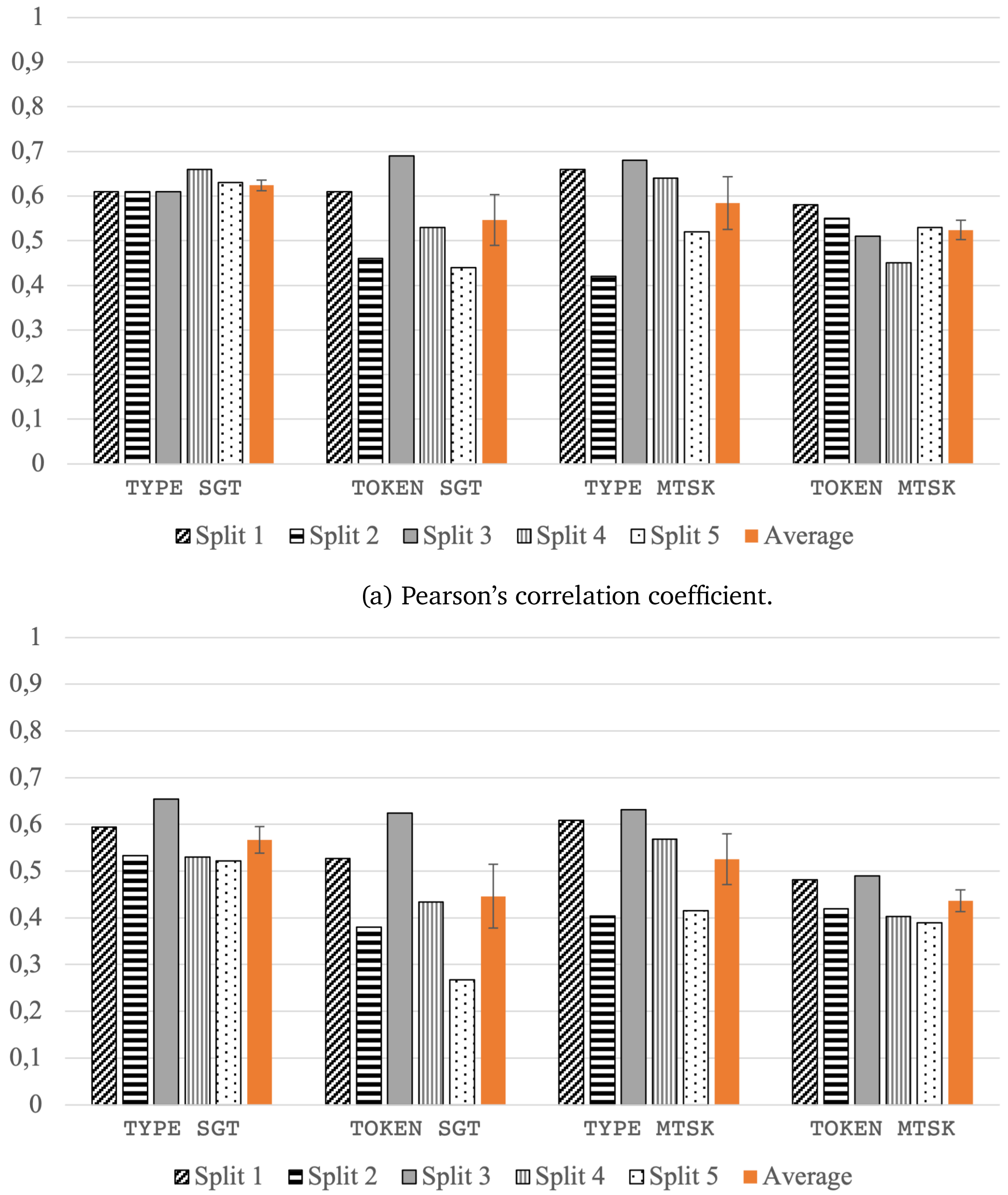
5.2.3 Influence of Distribution on Wug Experiment Correlations

For nonce verb inflection evaluation, we again computed the Pearson and Spearman correlation coefficients between A&H's human Wug Test production probabilities and the nonce verb predictions aggregated from the five runs on each split. For this second experiment, we also computed correlations on the irregular class. That is, correlations between the models' and human production probabilities for the irregular inflection of the nonce verbs. We computed the correlations for the regular and irregular class separately (cf. Corkery et al., 2019). This means that we computed the correlation between models' and human production probabilities for the regular nonce verb inflections, as well as for the irregular inflections. The correlations on the regular class represent how well the decision regular versus not regular is captured. The correlations on the irregular class also represent how well the model predicted the specific irregular inflection, e.g., *spling–splung* or *spling–splang*, or something else.

In Figures 5.2 and 5.3, we show the aggregate correlations per split. Again, the first bar represents the results from Experiment 1. We also include the average fold correlation and standard deviation. From these figures, it can be seen that $\texttt{TOKEN}_{\texttt{SGT}}$ obtained relatively unstable correlation coefficients on the regular and irregular class throughout the different splits, while $\texttt{TYPE}_{\texttt{SGT}}$ obtained relatively stable correlations. These observations make sense, given that different training-development distributions of verbs can drastically change the token frequency distribution of irregular inflection patterns in the training data. Some of these distributions may be more or less contributing to generalisation of past tense inflections in a human-like manner. When type frequency is used, this difference in frequency distribution of inflection patterns in the training data is less extreme, which explains the smaller differences in how the model generalises to the nonce verbs from the Wug Test.

The difference in variation of correlations between $\texttt{TOKEN}_{\texttt{SGT}}$ and $\texttt{TOKEN}_{\texttt{MTSK}}$, on the other hand, is less straightforward. $\texttt{TOKEN}_{\texttt{MTSK}}$ obtained more stable correlation throughout the different splits. For the token frequency model, it thus seems that the multi-task training setup had a stabilising effect. For the type frequency models, we see the opposite effect of multi-task training. In most cases, $\texttt{TYPE}_{\texttt{SGT}}$ demonstrated more stable correlations across the five splits than $\texttt{TYPE}_{\texttt{MTSK}}$.

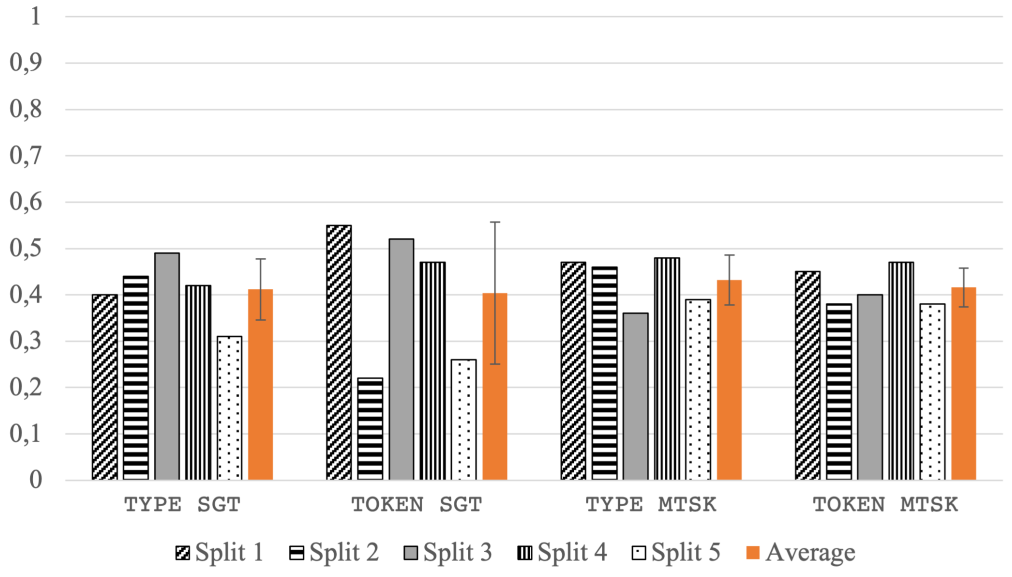
In conclusion, $\texttt{TYPE}_{\texttt{SGT}}$ and $\texttt{TOKEN}_{\texttt{MTSK}}$ have relatively stable correlations, regardless of different training-development set verb distributions, while $\texttt{TOKEN}_{\texttt{SGT}}$ and $\texttt{TYPE}_{\texttt{MTSK}}$ have relatively less stable correlations. These results demonstrate that, depending on the type of model, it is relevant to be mindful of which verbs and inflection patterns are maintained in the training data if—for instance—a set of real verbs is selected as a development set. In line with the conclusion from the previous section on real verb performance, we conclude from this that the aggregation of the results on all five splits gives the most fair comparison between the four models within our possibilities. Therefore, we aggregated the inflection predictions from all 25 runs. We present these results in more detail in the following Chapter (6).



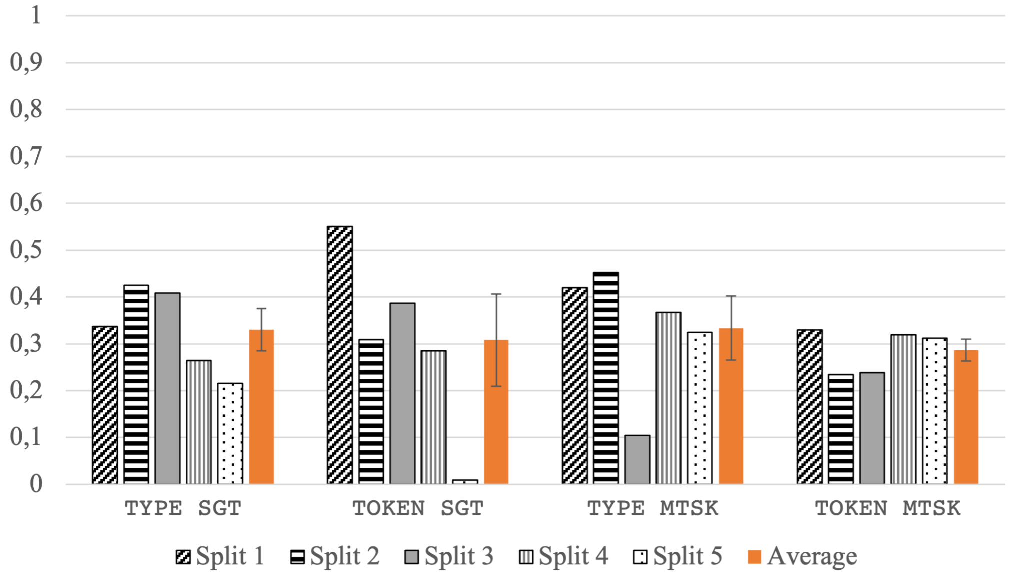
(a) Pearson's correlation coefficient.

(b) Spearman's correlation coefficient.

Figure 5.2: Correlation with human Wug Test data based on aggregate beam predictions for the regular class for each of the five training-development splits.



(a) Pearson's correlation coefficient.



(b) Spearman's correlation coefficient.

Figure 5.3: Correlation with human Wug Test data based on aggregate beam predictions for the irregular class for each of the five training-development splits.

Chapter 6

Overall Results

To gain a more comprehensive insight into the performance of the four models, we aggregated for each model the results from all five runs on each of the five splits.

6.1 Performance on Real Verbs

In this first part of the final analysis, we dive deeper into the model performance on real verbs. Like before, we discuss the models' average accuracies on the training and development set. Next, we also take a look at inflection error categories and auxiliary task performance.

6.1.1 Inflection Accuracy

In Table 6.1, we present the average accuracy on the training and development set of each of the four models. These results show that the differences in the average overall training and development accuracies are very small. On the training set, there is a slightly better overall performance of $\texttt{TOKEN}_{\texttt{SGT}}$ compared to the other models. On the development set overall, $\texttt{TOKEN}_{\texttt{MTSK}}$ performed slightly worse compared to the other three models. For the regular class, it can be noted that results are very similar for all models except $\texttt{TOKEN}_{\texttt{MTSK}}$, which performed slightly worse on both regular verbs from the training and development set. Nonetheless, these differences in accuracy are small.

	Training			Development		
	Overall	Irregular	Regular	Overall	Irregular	Regular
$\texttt{TYPE}_{\texttt{SGT}}$	97.11%	66.16%	99.20%	94.77%	34.96%	98.16%
$\texttt{TOKEN}_{\texttt{SGT}}$	99.51%	96.96%	99.71%	94.70%	45.31%	97.57%
$\texttt{TYPE}_{\texttt{MTSK}}$	97.62%	72.26%	99.54%	94.36%	34.62%	97.76%
$\texttt{TOKEN}_{\texttt{MTSK}}$	97.21%	91.80%	97.88%	92.77%	40.20%	95.79%

Table 6.1: Average accuracy on the training and test verbs of all four models (n=25).

A notable difference can be seen between token frequency and type frequency models on the irregular class. This is especially true for the training set accuracies. $\texttt{TOKEN}_{\texttt{SGT}}$ obtained the best accuracy on the irregular verbs of both the training and development set. With that, $\texttt{TOKEN}_{\texttt{SGT}}$ is the only model that performs near-perfect on irregular and regular training verbs, as well as regular verbs from the development set, like the models from previous studies (Corkery et al., 2019 and K&C). On the irregular class, TOKEN$_{\texttt{MTSK}}$ performs only somewhat worse than TOKEN$_{\texttt{SGT}}$. TYPE$_{\texttt{SGT}}$ shows the worst accuracy on the irregular verbs from the training set, with only ∼66% accuracy, and TYPE$_{\texttt{MTSK}}$ performed only somewhat better.

This difference between type and token frequency models on the irregular class performance makes sense, given that the token frequency models are trained on a set with proportionally more examples of irregular inflections compared to the type frequency models. It is also a result that we have seen throughout Experiment 1 and 2. In our early stopping criterion experiment (5.1.5), we saw that other type frequency models were able to obtain near-perfect accuracies on the irregular class from the training set as a result of more training epochs. However, the larger number of training epochs led to overfitting on the training examples in general, leading to a worsened fit with human behaviour on the nonce verbs. We discuss this finding further in the Discussion (Chapter 7).

All models obtained lower accuracies on the development set than on the training set, which is especially true for the irregular class. This makes sense because irregular verbs that are not seen during training can be hard or even impossible to generalise to. In the next section, we discuss this in more detail by categorizing the types of errors.

6.1.2 Inflection Error Analysis

We classified the inflection errors that we aggregated from all runs into different categories. We focussed on the following three classes: over-regularisations of irregular verbs, blending errors on irregular verbs, and *other* errors. In the case of over-regularisation, the model predicted a regular inflection for an irregular verb (e.g., *write–writed*). We classified inflections of irregular verbs as a blending when the inflection consisted of both the regular inflection as well as the correct irregular inflection (e.g., *write–wroted*). All other inflection errors were classified as *other*; think of errors such as predicting the wrong irregular form for an irregular verb (e.g., *sit–sit* or *sit–sut* instead of *sat*), predicting an irregular form for a regular verb (*mind–mound* instead of *minded*), or errors that have nothing to do with an inflection pattern but rather with the (phonetic) form in general (e.g., *execute–execued* instead of *executed*).

Figures 6.1 and 6.2 visualise for each model the distribution of all errors over the three classes for the training and development set, respectively. Along the Y-axis, the number of errors is indicated. Within the bars, the percentage per class is given. From Figure 6.1, we can conclude that the type frequency models over-regularise irregular verbs from the training set significantly more often, both proportionally and absolutely. Moreover, the majority of the type frequency models' errors can be explained by over-regularisation. This is in line with the fact that type frequency model accuracies on the irregular class are relatively poor, whereas their accuracies on the regular class are near-perfect.

On the development set, the difference in over-regularisation between the type and token frequency models is smaller than on the training set, as can be seen in Figure 6.2. This is mainly due to the fact that the token frequency models over-regularise irregular verbs from the development set more often than those from the training set. However, the type frequency models over-regularised still slightly more often than the token frequency models. This is in line with the performance of the four models on the development set, as presented in the previous section. On the irregular class, all four models perform relatively

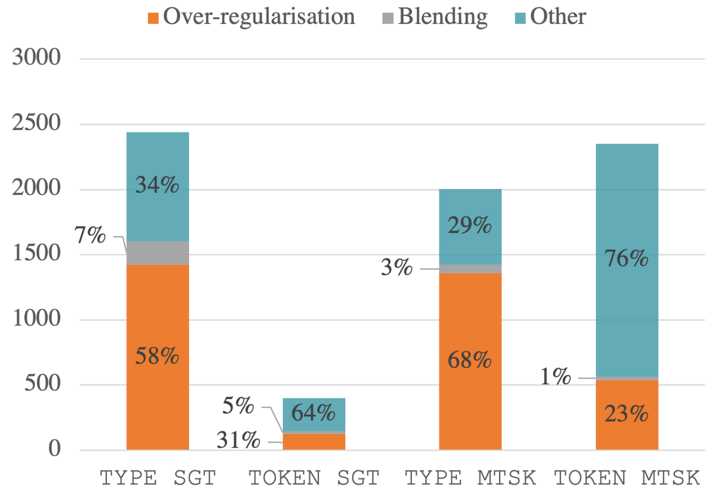


Figure 6.1: Classification of aggregate inflection errors on the training set (n=25).

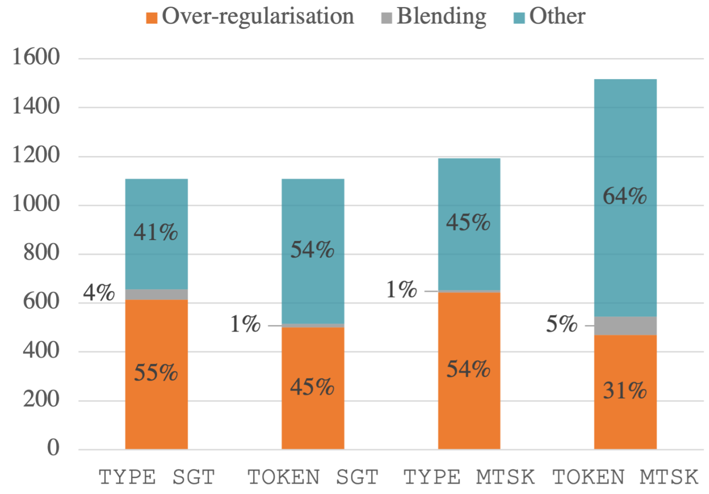


Figure 6.2: Classification of aggregate inflection errors on the development set (n=25).

poorly on the irregular verbs from the development set, although the type frequency models performed slightly more poorly. As explained, the over-regularisation of novel forms is in line with the human speaker's tendency to inflect novel forms regularly and is not necessarily seen as an undesired outcome. The limitation of the development set accuracy is that it does not fully provide insight into whether model generalisations align with human behaviour. Our evaluation of generalisation—the Wug Test—is presented in Section 6.2.3.

6.1.3 Label Accuracy Multi-task Models

As explained, the multi-task models did not only predict verb inflections, but also class labels (regular/irregular). In Table 6.2, the label prediction accuracies on the training and development set are given for $\texttt{TYPE}_{\texttt{MTSK}}$ and $\texttt{TOKEN}_{\texttt{MTSK}}$. From these results, we can conclude that the token frequency model outperformed the type frequency model on the irregular class of the training set, while they performed relatively similarly on the regular class. This is in line with the main task accuracies of $\texttt{TYPE}_{\texttt{MTSK}}$ and $\texttt{TOKEN}_{\texttt{MTSK}}$.

	Training			Development		
	Overall	Irregular	Regular	Overall	Irregular	Regular
$\texttt{TYPE}_{\texttt{MTSK}}$	76.70	33.76	84.97	81.41	25.90	84.48
$\texttt{TOKEN}_{\texttt{MTSK}}$	83.20	39.64	84.82	80.57	27.17	83.53

Table 6.2: Average accuracy on verb label prediction (n=25).

Furthermore, both models performed worse on this label prediction task compared to their main task of verb inflection predictions. Though this makes sense, given that this auxiliary task was assigned a lower weight in the loss optimization during training (0.3, as opposed to 0.7 for the main task), it is interesting that both models seem unable to predict verb labels as accurately as verb inflections. If we assume that the discrete binary distinction between regular and irregular verbs is an appropriate approach to past tense inflection—which aligns with the dual-route view of P&P—one would not expect this large difference in performance on the two tasks. We discuss this result further in the final discussion (Chapter 7).

6.2 Wug Experiment

In this second part of the chapter, we discuss the final results based on aggregating all nonce verb inflections and compute the correlations between these aggregate predictions and the A&H human Wug Test data. However, since we aggregate a larger set of results, we complement our analysis with a broader examination of the models' predictions first.

6.2.1 Beam Production Probabilities

We begin this analysis by discussing the beam production probabilities. For all four models, we found that the top 12 beam ranking probabilities are usually heavily skewed for high probabilities for the top 1 predictions. Furthermore, we were not able to find beams that contained more than one plausible form[1], which would always be the top 1 prediction. To illustrate this, the example below shows the top 5 from one beam prediction for the nonce verb nˈoʊld (*nold*). In this example, a model predicted the regular form nˈoʊldəd (*nolded*) with a high probability of > 0.99. We do not see any other plausible forms that could be expected here, such as nˈɛld (*neld*, as suggested by A&H). This is true for the whole top 12 in the beam prediction, but for simplicity, we present the top 5 here, as forms are also getting more bizarre towards the bottom of the beam.

1.	nˈoʊldəd	*nolded*	≈ 0.9990
2.	nˈoʊldədəd	*noldlded*	≈ 0.0003
3.	nˈoʊldoʊdəd	*noldloded*	< 0.0001
4.	nˈoʊldloʊdədəd	*noldlodeded*	< 0.0001
5.	nˈoʊldə	*nolde*	< 0.0001

Although less frequently, there were also instances where beam probabilities were less heavily skewed. *Bize* is an example of a nonce verb for which this happened relatively more

[1] A *plausible form* refers to a form that is morphologically legal and reasonable within the inflection rules.

frequently, and more often by the token frequency models. In the example below, $\texttt{TOKEN}_{\texttt{MTSK}}$ predicted the irregular inflection *boze* (which is also suggested by A&H as the most plausible irregular inflection). The probability assigned to this prediction is only $\sim$0.50. Again, all other predictions in the beam are implausible forms. We also see that the implausible forms combine the top form *boze* combined with the regular inflection *bized*, though the form *bized* itself does not appear in the top 12.

1.	bˈoz	*boze*	≈ 0.5041
2.	bˈoizd	*boized*	≈ 0.4436
3.	bˈoiɑt	*boiat*	≈ 0.0355
4.	bˈoiɑʒzd	*boia-uzed*	≈ 0.0031
5.	bˈoiɑʒouz	*boia-u-ouz*	≈ 0.0030

Our observations are in line with the results from Corkery et al. (2019), who found that the beam prediction of individual models usually contain only one plausible form at the top with a high probability, followed by implausible forms with low probabilities. They also found that aggregating the beam probabilities gave more insight into the overall preferences of the model as this washed out the unstable beam rankings of individual runs. This underlines the importance of focusing on the aggregate results as well as the consideration of other decoding strategies than beam search (see the Discussion, Chapter 7).

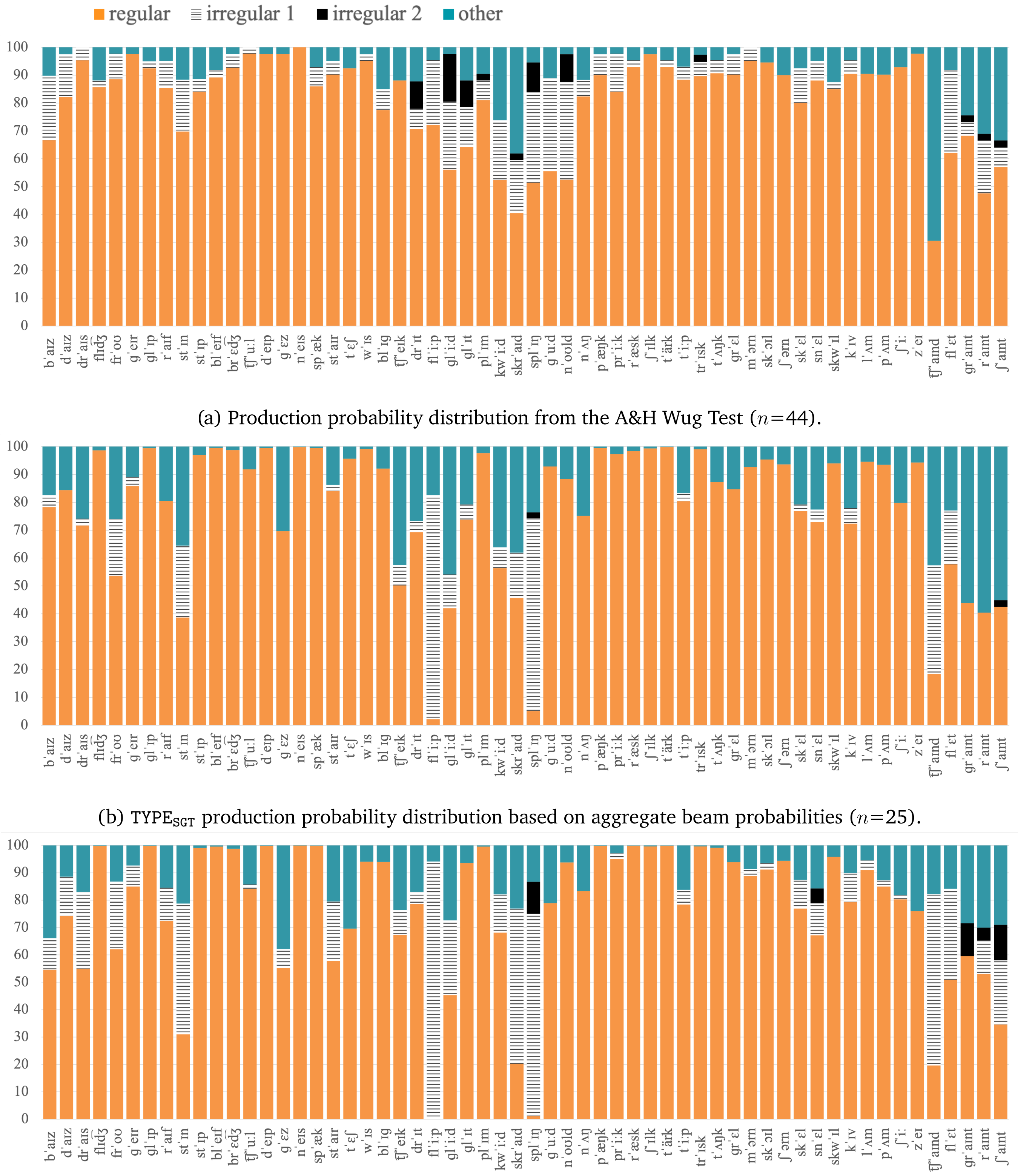
6.2.2 Production Probability Distributions

To gain insight into the overall predictions of our models, we aggregated and normalised the beam probabilities of the 25 model runs per category: *Regular, Irregular 1, Irregular 2,* and *Other* (as explained in the Methods, Section 4.4.3). These results are visualised in Figures 6.3b–6.3e. These distributions can be compared to the human speaker production probabilities from A&H's Wug Test, visualised in Figure 6.3a.

For all four models, the overall pattern corresponds to the human results presented in Figure 6.3a, which is that the regular inflection is most frequently used for most nonce verbs. In contrast with the individual beam probabilities, the aggregate beam results occasionally contain more than one plausible inflection. This is especially true for the token frequency models. For some nonce verbs, both the regular and irregular 1 inflections are predicted. In the case of the token frequency models, the irregular 2 form is present as well for a few nonce verbs, while this is minimally the case for the type frequency models.[2].

One difference between our models and human speaker data is that the models generally predicted higher probabilities for the *Other* class. However, we cannot compare the similarity between the predicted forms of this class since A&H did not publish these productions. Nevertheless, we do know from inspecting the beam rankings that the models predict implausible forms; sometimes as top 1 prediction, such as the blending of *bized* and *boze*: *bize–bozed*.

[2] Remind that A&H did only provide plausible irregular 2 forms for 11 out of 58 nonce verbs.

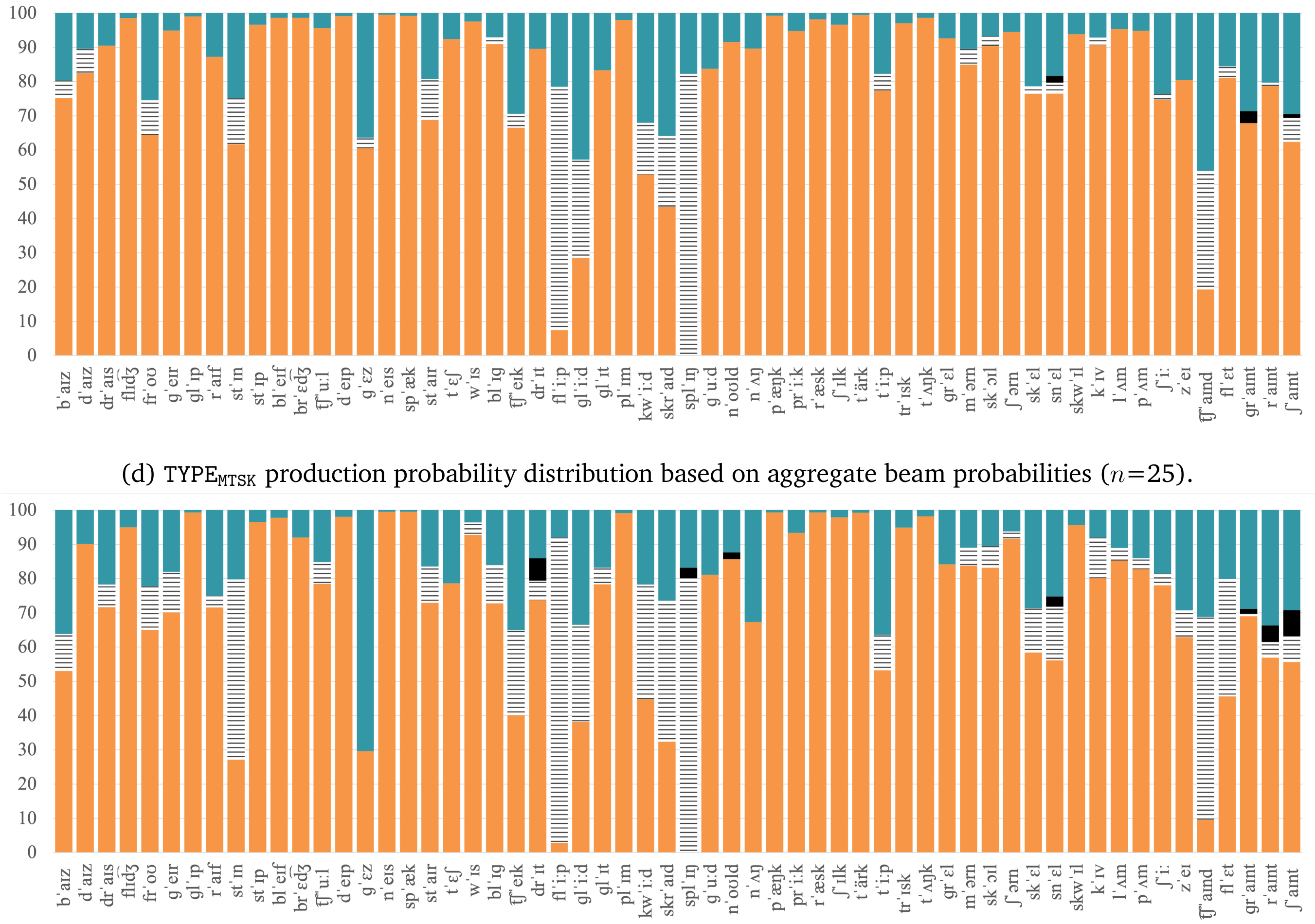


(a) Production probability distribution from the A&H Wug Test (n=44).

(b) TYPE$_{\text{SGT}}$ production probability distribution based on aggregate beam probabilities (n=25).

(c) TOKEN$_{\text{SGT}}$ production probability distribution based on aggregate beam probabilities (n=25).

Figure 6.3: Production probability distributions (Part 1 of 2).



(d) TYPE$_{\text{MTSK}}$ production probability distribution based on aggregate beam probabilities (n=25).

(e) TOKEN$_{\text{MTSK}}$ production probability distribution based on aggregate beam probabilities (n=25).

Figure 6.3: Production probability distributions (Part 2 of 2).

6.2.3 Wug Experiment Correlations

6.2.3.1 Production Probabilities

Like in Experiment 1 and 2, we compute the correlation between the human speaker data and the aggregate model predictions. Since we aggregated results from 25 model runs, we also computed production probabilities based on top 1 predictions. We counted for each nonce verb how often an inflection (*Regular/Irregular 1/Irregular 2/Other*) was the top prediction in the beam, and divided this by the total number of runs to compute a production probability distribution.[3] We argue that—when aggregating the results from a high number of runs—considering top 1 predictions is a realistic and human-like approach,

[3] Note that counting top 1 predictions based on only five runs, we only have five top 1 predictions per verb, which is not sufficient to compute a representative probability distribution with. This is why top 1 predictions were not used in the evaluations of Experiment 1 and 2.

given that A&H human production probabilities are an aggregation of one prediction per participant per verb as well.

In Table 6.3 we present for all four models the correlation coefficients between the aggregate results and the A&H human production probabilities. In Table 6.4 we present the *average* correlation between the human production probabilities and the 25 individual model runs.

	Regular				Irregular			
	Beam		Top 1		Beam		Top 1	
	r	ρ	r	ρ	r	ρ	r	ρ
TYPE$_{\mathtt{SGT}}$	**.69**	**.58**	**.63**	**.50**	.45	.40	.44	**.40**
TOKEN$_{\mathtt{SGT}}$	.63	.49	.62	.47	**.47**	.36	**.48**	.39
TYPE$_{\mathtt{MTSK}}$	.64	.55	.57	.47	.46	**.45**	.46	.35
TOKEN$_{\mathtt{MTSK}}$	.60	.47	.57	.47	.28	.42	**.48**	.29

Table 6.3: Pearson's (r) and Spearman's (ρ) correlation coefficients between model production probabilities (based on beam and top 1 predictions) and human production probabilities.

	Regular		Irregular	
	r	ρ	r	ρ
TYPE$_{\mathtt{SGT}}$	**.51**	**.51**	.34	.24
TOKEN$_{\mathtt{SGT}}$	.43	.41	.33	**.27**
TYPE$_{\mathtt{MTSK}}$	.50	.49	**.36**	.26
TOKEN$_{\mathtt{MTSK}}$	.41	.41	.33	.23

Table 6.4: Average Pearson's (r) and Spearman's (ρ) correlation coefficients between individual beam production probabilities and human production probabilities.

In Table 6.3 we generally see small differences between the correlations based on the aggregate results of the four models, especially for their top 1 predictions on the regular class. However, even if differences are small, for the regular inflection of the nonce verbs it is always TYPE$_{\mathtt{SGT}}$ with the highest correlation coefficients. TYPE$_{\mathtt{MTSK}}$ also scores higher or similar on the regular correlations compared to TOKEN$_{\mathtt{MTSK}}$. Finally, single task models have slightly stronger aggregate correlations than their multi-task versions on the regular class. For the average individual correlations, we can make similar observations, although the difference is somewhat more pronounced between type and token frequency models, and less pronounced between single and multi-task models.

In both the aggregate and average results, the correlations on the irregular class are usually relatively comparable between most models. The differences that can be found between the models, do not generally point to one of the models as doing consistently better or worse than others. This suggests that the four models generally share the struggle to capture human behaviour on the irregular class of nonce verb inflection.

Another observation that can be made is that all models generally have a higher correlation with the A&H data on the regular class than on the irregular class. This suggests that it is easier for our models to capture human behaviour in determining when to apply regular inflection, rather than predicting the specific type of irregular inflection. Furthermore, although the token frequency models have demonstrated significantly better performance on the irregular class of real verbs compared to the type frequency models, token frequency models do not consistently show a higher correlation on the irregular class.

In conclusion, there is a slightly better fit with human nonce verb inflections by type frequency models compared to token frequency models, and we usually found slightly worse correlations for the multi-task models compared to the single task models. These results have implications for the effect of token frequency and the validity of the dual-route approach. However, for a complete discussion of results we also must take into account results from the previously discussed sections and experiments. Therefore, our final discussion of the first and second research question is given in the Discussion (Chapter 7).

6.2.3.2 Ratings

In the second experiment of A&H's study, acceptability ratings of the nonce verb inflections were also elicited from the participants (n=24). On a scale of 1 to 7, participants indicated to what extent they found inflections if the nonce verbs acceptable. For completeness, we computed correlations between the model predictions and this rating data. Table 6.5 and 6.6 show the aggregate and average correlation coefficient of all models with respect to the rating experiment results of A&H.

	Regular				Irregular			
	Beam		Top 1		Beam		Top 1	
	r	ρ	r	ρ	r	ρ	r	ρ
TYPE$_{\text{SGT}}$	**.71**	**.65**	**.65**	**.60**	.40	**.57**	.40	**.50**
TOKEN$_{\text{SGT}}$	.63	.54	.62	.51	.35	.32	.34	.31
TYPE$_{\text{MTSK}}$	.69	.60	.62	.51	.38	.44	.38	.37
TOKEN$_{\text{MTSK}}$	.61	.49	.59	.41	**.43**	.38	**.44**	.39

Table 6.5: Pearson's (r) and Spearman's (ρ) correlation coefficients between model production probabilities (based on beam and top 1 predictions) and human rating data.

	Regular		Irregular	
	r	ρ	r	ρ
TYPE$_{\text{SGT}}$	.54	**.55**	**.31**	.24
TOKEN$_{\text{SGT}}$	.44	.44	.24	.23
TYPE$_{\text{MTSK}}$	**.55**	.53	.29	.26
TOKEN$_{\text{MTSK}}$	.42	.43	.30	**.28**

Table 6.6: Average Pearson's (r) and Spearman's (ρ) correlation coefficients between individual beam production probabilities and human rating data.

As can be seen from the tables, the human rating data led to relatively similar results as the human production probabilities. However, the preference of type frequency models over token frequency models on the regular class is somewhat stronger here. Like before, the differences are less pronounced for the correlations based on the models' top 1 predictions. Finally, compared to the correlations based on the human production data, we see here slightly stronger correlations for the regular class.

6.2.3.3 Nonce Verb Label Prediction

To evaluate the multi-task models' label predictions on nonce verbs, we calculated the correlation between the human production probabilities from A&H's Wug Test and the model predictions. For each nonce verb, the multi-task models predicted a probability for the regular class label. We aggregated these probabilities across all runs, similar to how we did for the beam probabilities, and computed the correlation with the human data. Additionally, we calculated this correlation by counting how often the regular class was predicted for each verb (i.e., $p > .50$ for regular class label, referred to as Top 1) and divided this by the number of runs. In Table 6.7, these correlations are shown. We also repeat the results from the multi-task model inflection predictions for comparison.

	Labels			
	Probability		Top 1	
	r	ρ	r	ρ
TYPE$_{\text{MTSK}}$ labels	**.53**	**.38**	**.56**	**.39**
TOKEN$_{\text{MTSK}}$ labels	.51	.36	.49	.34
TYPE$_{\text{MTSK}}$ inflections	.64	.55	.57	.47
TOKEN$_{\text{MTSK}}$ inflections	.60	.47	.57	.42

Table 6.7: Correlations between human production probabilities and aggregated label probabilities and top 1 predictions; and repeated correlations between human production probabilities and beam and top 1 inflection predictions. Pearson (r) and Spearman (ρ) on the regular class.

In line with the main task results, the type frequency model somewhat outperforms the token frequency model. However, the correlations based on label predictions are weaker compared to those based on the inflection predictions of the main task. This difference can be explained by the fact that the class label prediction and inflection prediction do not always align. For TYPE$_{\text{MTSK}}$, the label and inflection predictions agree on 88% of the nonce verbs of the Wug Test. It was 14 times the case that a regular label and irregular inflection were predicted for the same form, while it was 159 times the case that an irregular label and regular inflection were predicted for the same form. For TOKEN$_{\text{MTSK}}$, this is the case for 90% nonce verbs. It predicted 48 times a regular label while predicting an irregular inflection, and 94 times an irregular label while predicting a regular inflection.

We conclude from these results that label prediction and inflection prediction often agree, but not always. In case of disagreement between the two, the models tend to label verbs more often as irregular than the inflection prediction does. Like with the accuracy on real verbs, the fit of the models' label predictions with human data is worse compared to their inflection predictions. As mentioned before, this is not the expected result if it is accurate to make a discrete distinction between regular and irregular verbs and assume that they are separately processed, as suggested by the dual-route view of P&P. We discuss this further in the Discussion Chapter (7) in light of the second research question.

6.2.4 Regular Inflection Tendency

Finally, we take a look at the models' tendency to predict the regular inflection on novel inputs. We have seen in Section 6.1.2 and 6.2.2 that there is a global tendency of models—like humans—to produce a regular inflection on novel forms. Moreover, over-regularisations were more often made by type frequency models on real verbs. Though this contributed to a worse accuracy on irregular verbs compared to token frequency models, they also slightly outperformed token frequency models on correlations with human data on the Wug Experiment. To discover whether a stronger tendency for the regular inflection on the nonce verbs co-occurs with higher correlations, we computed different measures of regular inflection tendency. In Table 6.8, we show for each model their average production probability for the regular inflection, the percentage of nonce verbs where the regular inflection probability is $\geq 50\%$, and the percentage of nonce verbs where the regular inflection probability has the largest probability compared to the other inflection classes. These results are presented in Table 6.8.

	Average probability	$p > 50\%$	Largest probability
Human data	81%	95% (55/58 verbs)	98% (57/58 verbs)
TYPE$_{\text{SGT}}$	78%	84% (49/58 verbs)	88% (51/58 verbs)
TOKEN$_{\text{SGT}}$	76%	88% (51/58 verbs)	91% (53/58 verbs)
TYPE$_{\text{MTSK}}$	81%	92% (53/58 verbs)	93% (54/58 verbs)
TOKEN$_{\text{MTSK}}$	72%	83% (48/58 verbs)	90% (52/58 verbs)

Table 6.8: Different measures of regular inflection tendency on the Wug Test.

First of all, TYPE$_{\text{MTSK}}$ has the highest regular inflection tendency. Depending on the measure, TYPE$_{\text{SGT}}$ and TOKEN$_{\text{SGT}}$ follow. This pattern does not perfectly fit the finding that TYPE$_{\text{SGT}}$ has the best overall fit with human behaviour. However, the differences between TYPE$_{\text{SGT}}$, TOKEN$_{\text{SGT}}$ and TYPE$_{\text{MTSK}}$ were usually small and sometimes even negligible. Moreover, TOKEN$_{\text{MTSK}}$ having the significantly weakest regular inflection tendency aligns with having the weakest correlations with the human data on the regular class. We conclude from the results that a stronger tendency for regular inflection does not precisely go hand-in-hand with a better fit with human behaviour on novel input. This indicates that, as can be expected, there are more subtle intricacies that play a role in a good fit with human behaviour on novel verbs.

Something else that does seem to exactly match the regular inflection tendency of models is the number of training epochs. TYPE$_{\text{MTSK}}$ trained for the highest number of average epochs (50), while TOKEN$_{\text{MTSK}}$ trained for the lowest average number of epochs (17), leaving TYPE$_{\text{SGT}}$ and TOKEN$_{\text{SGT}}$ in the middle (with on average 28 and 27 epochs respectively). This suggests that using more training epochs results in an overall stronger tendency to use the regular inflection on novel input.

Chapter 7

Discussion

In this chapter, we discuss the results of this thesis in light of the posed research questions in Section 1.1. Finally, we propose several directions for future research.

7.1 The Effect of Token Frequency

The first sub-question that we posed is: *Does training an ED model on token frequency instead of type frequency lead to more similar real and nonce verbs inflections to those of human speakers?* To find an answer to this question, we trained ED models similar to the ones used in Corkery et al. (2019) and K&C on type frequency and token frequency. In case of type frequency ($\texttt{TYPE}_{\texttt{SGT}}$), the training data of the model contained each example once. This is in line with the input data of previous studies (A&H; Corkery et al., 2019; K&C). In the case of token frequency ($\texttt{TOKEN}_{\texttt{SGT}}$), the training data was augmented with token frequency distribution. This was done by adding verb examples multiple times in the training data, proportional to how often its past tense form occurs in the CELEX corpus. Since for the second sub-question we investigated the effect of a multi-task training setup, we also made a type and a token frequency model in this setting: $\texttt{TYPE}_{\texttt{MTSK}}$ and $\texttt{TOKEN}_{\texttt{MTSK}}$. Hence, we compared two pairs of type and token frequency models.

Aggregating the results from Experiment 1 and Experiment 2, we first compared the models to each other on average accuracy on the training and development set, containing only real verbs. Though the regular class performance was relatively similar between the models, the token frequency models strongly outperformed the type frequency models on the irregular class from the training set. Token frequency models also outperformed the type frequency models on the irregular verbs from the development set, but this difference was smaller. In Experiment 2, we noted that type models had much higher standard deviations on the irregular class accuracy as well. As pointed out throughout the experiments, the difference in accuracy could be explained by the fact that the token frequency models have proportionally more exposure to irregular examples compared to the type frequency models.

Using the Wug Test as the test of generalisation, we compared the models' nonce verb inflections by computing the correlations between the model inflection predictions and the A&H human Wug Test data. We found that the type frequency models generally obtain higher correlations on the regular class compared to the token models. However, it is important to mention that the token frequency models obtained only slightly worse correlations based on the aggregate beam results, and we found no notable difference between correlations based on the aggregate top 1 predictions.

Insights from Experiment 1 generally led to a selection of models with higher nonce verb correlations, also compared to those of previous studies (Corkery et al., 2019; K&C). However, we also observed that the irregular class accuracy of the type frequency models significantly dropped from near-perfect accuracy on the training set to 65%–75%. Although we prioritised capturing human behaviour on novel inputs, as this evaluates the ability of models to generalise, this came with the cost of performing worse on real irregular verbs for the type frequency models. This is in line with the findings of Corkery et al. (2019), who also explored using fewer training epochs for a model similar to $\texttt{TYPE}_{\texttt{SGT}}$, and found indeed higher Wug Test correlations but lower training performance, especially on the irregular class.

In contrast, the token frequency models did not face this issue, as they achieved near-perfect accuracy on the training verbs and relatively similar aggregate correlations with human Wug Test data. As mentioned in the Theoretical Background and Introduction, the token frequency distribution of verbs is also more similar to the linguistic input of human language learning. In conclusion, we consider token frequency to be a valuable aspect that makes ED models a more representative cognitive model of English past tense inflection. This finding is particularly noteworthy, as previous studies on past tense inflection consistently relied on type frequency rather than token frequency (e.g., A&H; Corkery et al., 2019; K&C). Our results also contribute to our understanding of the cognitive processes underlying English past tense inflection, suggesting that token frequency might play a more important role than previously assumed.

7.2 The Effect of Dual Route Multi-task Training

The second sub-question that we posed is: *Does training an ED model on an auxiliary task of distinguishing regular and irregular verbs lead to more similar inflections of real and nonce verbs to those of human speakers?* To investigate this question, we experimented with models that were not only trained on the main task of predicting verb inflections, but also on an auxiliary task to predict the verb class (regular or irregular). The objective of this multi-task training is to reflect in the task of the model the P&P dual-route proposition that regular and irregular verbs are distinctly processed in English past tense inflection. To find an answer to this question and gain more insight into the plausibility of the dual-route approach, we compared the multi-task trained models ($\texttt{TYPE}_{\texttt{MTSK}}$ and $\texttt{TOKEN}_{\texttt{MTSK}}$) to their single task versions ($\texttt{TYPE}_{\texttt{SGT}}$ and $\texttt{TOKEN}_{\texttt{SGT}}$).

In Experiment 1, different weighted sums of the overall loss were investigated. Results demonstrated that a weight of 0.7 to the main task of inflection prediction and 0.3 to the auxiliary task of classification led to the strongest fit with human nonce verb inflections. This was compared to the two other options where the weighted sum was (0.8, 0.2) and (0.6, 0.4) to the main and auxiliary tasks, respectively. The model with the lowest weight assigned to the auxiliary task (0.2) is obtained stronger correlations with human nonce vebr data than one with a heavier weight for the auxiliary task (0.3). This would suggest that the auxiliary task adds something useful. However, the results also demonstrated that the weighted sum of (0.7, 0.3) led to the label performance being significantly worse than in the other settings.

Comparing the aggregate results of the selected models, we found that the multi-task models usually perform either worse or similar on the training and development verbs. The only improvement observed was $\texttt{TYPE}_{\texttt{MTSK}}$ performing slightly better on irregular verbs from the training set compared to $\texttt{TYPE}_{\texttt{SGT}}$. This is also true for the Wug Test correlations: overall, the single task models yielded either similar and otherwise usually slightly stronger overall correlations than their multi-task versions. This is regardless of using type frequency or token frequency. This means that inflections on real and nonce verbs cannot be considered more similar to those of humans when the model is additionally focusing on the discrete distinction between regular and irregular by means of an auxiliary task.

Another relevant finding was the difference between the main task and auxiliary task performance. Overall, the label predictions were less accurate on the real verbs and also yielded lower correlations with human Wug Test data, compared to the inflection predictions. Label predictions from the auxiliary task and inflection predictions from the main task matched about 90% of the times for both models on nonce verb predictions. This was largely due to the models predicting an irregular label for a nonce verb that they predicted a regular past tense form for. Interestingly, the model has trouble matching its inflection and label predictions to some extent. Of course, the auxiliary task was assigned a lower weight, which explains the better performance on the main task compared to the performance on the auxiliary task. However, if this binary classification of verbs is a completely accurate way to model past tense inflection, one may not expect this difference between the main and auxiliary task predictions.

In conclusion, the relatively small performance differences between the single-task and multi-task models on real and nonce verbs suggest that adding the auxiliary task may not significantly alter the models' representations. It is possible that the auxiliary task, designed to reflect the dual-route idea, compels the models to learn abstract representations similar to those learned without the auxiliary task. Such a finding would be consistent with P&P's dual-route view, as well as with Ma and Gao (2022) their findings that their transformer model captures an abstract distinction between regular and irregular verbs. However, we also observed mismatches between label predictions and inflection predictions, which challenge this interpretation. Therefore, we consider it a feasible explanation that performance was either similar or slightly worse in the multi-task training setup, because the binary distinction of verbs is not exactly accurate and beneficial for human-like inflection generalisations, and the auxiliary task may not have entirely resulted in the dual-route-like representation despite its intended design.

Future research could further investigate this by using a similar multi-task training setup but experimenting with three classes instead of two, *regular, semi-regular,* and *irregular*, to investigate whether this is a more optimal classification of verbs. Another option would be to interpret the obtained representations of the model, as briefly discussed in this chapter (Section 7.5.3).

7.3 Model Configuration Choices for Encoder-Decoders as Models of Inflection

In Experiment 1 (Section 5.1), we explored different model configuration choices to answer the question: *Which data and model configuration choices improve an ED model's ability to predict more similar inflections of real and nonce verbs to those of human speakers*. As explained, we adopted a similar model to the one implemented by Corkery et al. (2019) and K&C. However, exploration and investigation of different model configuration choices for the specific task of this model had not been conducted before. Our experiment also allowed for a model selection to make the most fair comparison between the four models that we investigated in this thesis. The final selection of models demonstrated an improvement on the correlations obtained in the previous study of Corkery et al. (2019), where similar evaluation led to regular class Wug Test correlations of .30/.45 (Pearson and Spearman, respectively) instead of .69/.58 (TYPE$_{\text{SGT}}$). Below, we discuss the most important observations in Experiment 1 that led to these improvements.

7.3.1 Representativeness of the Data

The first steps of Experiment 1 considered different versions of the dataset. Most importantly, we discovered that a set of highly frequent and irregular verbs was omitted from the A&H dataset that was also used by Corkery et al. (2019) and K&C. The results showed that nonce verb inflections were more similar to those of humans when these verbs were actually included in the data.

We drew a similar conclusion regarding the inclusion of more than one correct inflection for some verbs (e.g., *dive–dove/dived*). Although previous studies have also chosen to include these in the data, it is not straightforward to assume that this is the best option without any empirical exploration. This is because, for instance, in cases homophones such as *wring/ring* (*wrung/rang*), human speakers usually are provided with semantic context that allows for disambiguation between these verbs, whereas the model is not.

Similar to our conclusion with regard to the dataset augmentation with token frequency, these findings suggest that the above alternations to the dataset both made the data more representative of the linguistic input of human language learning as well as improved all four models' fit with human behaviour on nonce verbs.

7.3.2 Preventing Overfitting

Another main finding from Experiment 1 is that mitigating the suspected problem of overfitting on the real verb training data improved the match between model generalisations to novel inputs and human data. Overfitting on the real verb data was mitigated by using an early-stopping mechanism with an appropriate patience threshold parameter, as well as an appropriate selection of hyperparameter settings, including network depth and size, dropout, batch size and learning rate. However, for the four models, it differed which settings resulted in optimal generalisations to nonce forms.

A consequence of this, however, is that less overfitting of the type frequency models led to a significantly weaker performance on real irregular verbs, compared to the near-perfect accuracies also found by the previous studies (Corkery et al., 2019; K&C). Within the scope of Experiment 1, we have not been able to discover a type frequency model that is able to perform well on real verbs and obtain equally high correlations with human data. When the training data was augmented with the token frequency distribution, the models performed with near-perfect accuracy on real verbs and correlated almost as strongly with human data as TYPE$_{\text{SGT}}$. Therefore, we suggest that the described improvements to prevent overfitting go hand-in-hand with augmenting the training data with the token frequency distribution.

In conclusion, our results show that both model configuration choices regarding model architecture, training and hyperparameters—especially those influencing the risk of overfitting—improved a model's capacity to capture human behaviour on nonce verbs.

7.4 Distribution of Verbs

A downside of using an early-stopping criterion is that we had to split the real verb data into a training and development set to be able to track convergence on the development verbs. As explained in this thesis (Sections 3.4.4, 5.2), the problem with this is that it is not a straightforward task to find a training-development split for verb data that suits this task: a training set containing all the relevant information to generalise in a human-like manner to the nonce verbs, as well as a development set that is a representative set of the training data. Though for Experiment 1, we relied on one random split, following the conventional approach in the field of machine learning, we further investigated the effect of this with our fourth sub-question: *Does the distribution of real verbs over a training and development set influence an ED model's fit with human behaviour on real and nonce verb inflections?*

Experiment 2 mainly revealed that different distributions of verbs into training and development parts can lead to differences in performance on real verbs as well as a model's fit with A&H human Wug Test data, but this differed between the models. As mentioned before, the performance of the type frequency models is weaker and less stable compared to the more robust token frequency models. On the nonce verbs, the correlation strength with human data also depended on the particular set of verbs the models are trained on, in particular for $\texttt{TOKEN}_{\texttt{SGT}}$ and $\texttt{TYPE}_{\texttt{MTSK}}$).

This raises the issue of identifying an optimal split to ensure the training and development sets are as representative as possible. Unlike K&C, who relied on a single random data split, we chose to aggregate results across multiple splits, as this approach was the most suitable within the scope of this thesis to handle this issue. Future research could also look further into which distribution of verbs results in the most representative split so that the data may divided into an optimal training-development split, and an early-stopping criterion can still be used to help prevent overfitting on the training verbs. Another option for future research would be to use all verbs as training data. This eliminates the problem of finding an ideal distribution of verbs and replicates the task of language acquisition most accurately (cf. Corkery et al., 2019). However, an appropriate fixed number of epochs should be used in order to make sure that sufficient training epochs are used without overfitting.

7.5 Additional Directions for Future Research

In this final section of the Discussion, we propose several directions for future research based on the findings and limitations of this thesis.

7.5.1 Decoding Strategy

Inspecting the aggregate beam results revealed that the beam search predictions contains at most only one plausible form as the top 1 prediction. All other predictions appeared to be forms that are unnatural to use in human language. Usually, the probabilities were also heavily skewed for the plausible top 1 prediction. A possible way to overcome limitations of the beam search could be using other decoding strategies instead of beam search. Examples of this are nucleus sampling (Holtzman et al., 2019) or top-k sampling(Fan et al., 2018; Radford et al., 2019). These techniques could help improve observed problems such as repetition (see also examples in Section 6.2.1).

In this thesis, however, we have focused on another way to overcome the beam search algorithm limitations, which is by focusing on the aggregate top 1 predictions. When aggregating model predictions across a sufficient number of runs, top 1 predictions could be considered similar to the A&H Wug Experiment, where they elicited from each of the 41 participants one past tense form for each nonce verb. Hence, this not only eliminates the beam search limitations but also makes the experimentation more similar to that of human speakers.

7.5.2 Inflection Trends and Patterns

Another observation from the overall results is that the correlations between the models' production probabilities and those of human speakers were generally stronger on the regular class compared to those on the irregular class. This is true for all four models, as well as the models from previous studies (A&H; Corkery et al., 2019; K&C). As mentioned in the previous chapter, this implies that it is more straightforward whether humans would inflect a verb regular or irregular, compared to the decision of *which* irregular pattern to apply.

Furthermore, we found that the tendency of models to inflect novel input regularly could only partially explain the strength of the correlations. Of course, the regular inflection being highly productive must be captured in order to mimic human inflections on novel inputs. However, more subtle and complex intricacies underlie the decisions of humans to inflect novel forms.

A limitation of the models related to this is the fact that the models made past tense inflection errors that are a blend of both regular and irregular inflection. We saw this happen on the training, development and nonce verbs (for instance, *bize–bozed*). This type of error is also noted by P&P as a limitation of R&M's connectionist model. K&C noted that their model (though obtaining significantly lower correlations than ours) overcame this problem and did not find this error type. These errors pose a limitation of our models' fit with human behaviour, as these are not human-like errors (Pinker, 1999).

Future research could focus on a more detailed understanding of these behavioural nuances and generalisation patterns in neural ED models.

7.5.3 Interpretation of Encodings

A general limitation, often pointed out, is that neural networks are 'black boxes' with limited insight into what they learn (McCloskey, 1991). This is especially relevant in cognitive modelling, where the goal is to gain more insight into human behaviour and the underlying cognition. However, future research could address this limitation by applying more advanced analytical methods. Corkery et al. (2019) did this by making a t-SNE (t-distributed Stochastic Neighbour Embedding) visualization of the verb encodings (Van der Maaten and Hinton, 2008). The multi-dimensional encodings of the model are mapped onto a two-dimensional space, which visualises which verb encodings are close to each other. This can be useful to understand where specific inflection decisions of the models come from.

Another way for future studies to interpret the model encodings could be by implementing probing classifiers (Belinkov, 2022). For instance, a binary classifier could be trained to predict whether a model's encoding represents a regular or irregular verb. The purpose of this probing classifier is to reveal the specific properties encoded by the model. In this setup, the classifier's ability to accurately distinguish between regular and irregular verb forms would indicate whether the model abstractly captures this distinction in its representations. This approach could provide valuable insights into the validity of the dual-route theory, complementing the findings from our multi-task training.

7.5.4 Other Languages

Finally, we want to highlight the importance of future research to focus on other kinds of morphological inflection and languages besides English past tense inflection. One of the studies that have already done this are McCurdy et al. (2020), who focused on German plural inflection of nouns, and Yang et al. (2023) (using the UniMorph dataset McCarthy et al., 2020) who focused on past tense inflection in English, Dutch and German. Different cases of inflection in different languages pose different challenges. For example, in German plural inflection, there are multiple regular rules, and there is no majority class like the regular past tense inflection in English. Looking into different cases and languages provides more insight into the general capacity of these neural models as cognitive models of morphological inflection.

Chapter 8

Conclusion

In this thesis, we investigated our main research question of whether neural models, specifically EDs, have the potential to cognitively model English past tense inflection. Despite the strong scepticism following P&P's rebuttal of R&M's connectionist model in of past tense inflection in 1986, recent work (a.o., Corkery et al., 2019 and K&C) demonstrated that contemporary, advanced neural models overcome many theoretical and empirical limitations of R&M's model. However, Corkery et al. (2019) also showed that their ED model was not able to precisely capture human behaviour on nonce forms.

In this thesis, we have followed up on the studies of Corkery et al. (2019) and K&C by investigating the effect of different data, models and training configurations. Our results demonstrated that the representativeness of the data influenced a model's ability to generalise in a human-like manner to novel inputs. Making the data more complete, hence more similar to the linguistic input of human language learning, improved its fit with human behaviour on nonce forms. The representativeness of the data must also be taken into account if verbs are distributed into a training and development set, as our results demonstrated that this can influence model performance as well.

Furthermore, our results confirmed that the suspected overfitting of the previous studies' models could be prevented with the use of an early stopping mechanism and appropriate hyperparameter settings. This led to a substantial improvement in correlations between model predictions and human Wug Test data. Additionally, models achieved near-perfect performance on regular verbs from the training and development sets, comparable to that reported by Corkery et al. (2019). However, adjustments such as fewer training epochs also led to a decrease in model performance on irregular real verbs in the training set, which was observed by Corkery et al. (2019) as well. This is undesirable for a cognitive model of past tense inflection, as a model preferably also captures that speakers perform near-perfect on their acquired set of verbs.

A key finding of this thesis is that augmenting the training data with the token frequency distribution of verbs led to a substantial increase in accuracy on irregular verbs in the training set ($\simeq$96% compared to the type frequency model's $\simeq$66%). Additionally, correlations were almost equally strong as those obtained with type frequency when aggregating predictions across multiple model instances (cf. Corkery et al., 2019). Our results show that augmenting the training data with token frequency does not only make the input more similar to those of humans, but also is valuable in increasing the overall fit with human behaviour. This is a relevant finding, given that previous studies modelling past tense inflection consistently chose to only reflect type frequency in the training data (a.o., A&H; Corkery et al., 2019; K&C).

We also investigated whether a multi-task training setup additionally focusing on the discrete distinction between regular and irregular verbs—intended to represent ideas of P&P's dual-route approach—improves a model's fit with human behaviour on real and nonce verbs. Our results did not demonstrate that this was the case, as only relatively small differences in real and nonce verb performance were found, often to the disadvantage of the multi-task setup. Based on our findings, neither definite confirmation nor rejection of the dual-route approach could be determined. Nonetheless, our experiments demonstrate that contemporary neural models can be a meaningful way to experiment and gain insight into long debated topics such as English past tense inflection. These models have a large learning capacity and the flexibility to integrate and empirically experiment with different theories, even if they initially opposed the neural modelling approach. Future research could further investigate the plausibility of approaches such as the dual-route view in a similar way.

Although our results demonstrate improvements in generalising to novel inputs in a more human-like way compared to previous studies, our results suggest that ED models still face the challenge to capture human behaviour on a detailed level, such as on the irregular inflection of nonce verbs and occasionally predicting unrealistic forms (e.g. blendings, *dive–doved*). Nevertheless, our findings suggest that there are promising avenues to further improve the alignment of these models with human behaviour.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., ... Zheng, X. (2015). Tensorflow: Large-scale machine learning on heterogeneous systems [Software]. https://www.tensorflow.org/

Ahlberg, M., Forsberg, M., & Hulden, M. (2015). Paradigm classification in supervised learning of morphology. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1024–1029.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition*, *90*(2), 119–161.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). CELEX2 LDC96L14 [Web download]. Linguistic Data Consortium.

Baayen, R. H. (2009). Corpus linguistics in morphology: Morphological productivity. In A. Lüdeling & M. Kyto (Eds.), *Corpus linguistics: An international handbook* (pp. 899–919). De Gruyter Mouton. https://doi.org/10.1515/9783110213881.2.899

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bauer, L. (1983). *English word-formation*. Cambridge University Press.

Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, *48*(1), 207–219.

Berko, J. (1958). The child's learning of english morphology. *Word*, *14*(2-3), 150–177.

Beser, D. (2021). Falling through the gaps: Neural architectures as models of morphological rule learning. *arXiv preprint arXiv:2105.03710*.

Blything, R. P., Ambridge, B., & Lieven, E. V. (2018). Children's acquisition of the english past-tense: Evidence for a single-route account from novel verb production data. *Cognitive Science*, *42*, 621–639.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and cognitive processes*, *10*(5), 425–455.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Corkery, M., Matusevych, Y., & Goldwater, S. (2019). Are we there yet? encoder-decoder neural networks as cognitive models of english past tense inflection. *arXiv preprint arXiv:1906.01280*.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., & Hulden, M. (2016). The sigmorphon 2016 shared task—morphological reinflection. *Proceedings of the 14th SIGMORPHON workshop on computational research in phonetics, phonology, and morphology*, 10–22.

Durrett, G., & DeNero, J. (2013). Supervised learning of complete morphological paradigms. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1185–1195.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Fan, A., Lewis, M., & Dauphin, Y. (2018, July). Hierarchical neural story generation. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 889–898). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1082

Faruqui, M., Tsvetkov, Y., Neubig, G., & Dyer, C. (2015). Morphological inflection generation using character sequence to sequence learning. *arXiv preprint arXiv:1512.06110*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Kann, K., & Schütze, H. (2016). Med: The lmu system for the sigmorphon 2016 shared task on morphological reinflection. *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 62–70.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirov, C. (2023). Revisitpinkerandprince: Experiment 1 wugs [Accessed: 2023-01-31]. https://github.com/ckirov/RevisitPinkerAndPrince/tree/master/experiment_1_wugs

Kirov, C., & Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, *6*, 651–665.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day american english*. Dartmouth Publishing Group.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Ma, X., & Gao, L. (2022). How do we get there? evaluating transformer neural networks as cognitive models for english past tense inflection. *arXiv preprint arXiv:2210.09167*.

Marcus, G. (1998). Rethinking eliminative connectionism. *Cognitive psychology*, *37*(3), 243–282.

Marcus, G. (2020). The next decade in ai: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., Arkhangelskiy, T., Krizhanovsky, N., Krizhanovsky, A., Klyachko, E., Sorokin, A., Mansfield, J., Ernštreits, V., Pinter, Y., Jacobs, C. L., . . . Yarowsky, D. (2020, May). UniMorph 3.0: Universal Morphology. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the* *twelfth language resources and evaluation conference* (pp. 3922–3931). European Language Resources Association. https://aclanthology.org/2020.lrec-1.483

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological science*, *2*(6), 387–395.

McCurdy, K., Goldwater, S., & Lopez, A. (2020). Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for german plurals. *arXiv preprint arXiv:2005.08826*.

Nicolai, G., Cherry, C., & Kondrak, G. (2015). Inflection generation as discriminative string transduction. *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, 922–931.

Osborne, J. W. (2002). Normalizing data transformations. ERIC digest. https://eric.ed.gov/?id=ED470204

Pierrehumbert, J. (2001). Stochastic phonology. *Glot international*, *5*(6), 195–207.

Pinker, S. (1999). *Words and rules: The ingredients of language*. Basic Books.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73–193.

Plag, I. (2003). *Word-formation in english*. Cambridge University Press.

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, *8*(1), 1–56.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of english verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 216–271). MIT Press.

Schultink, H. (1961). Produktiviteit als morfologisch fenomeen. *Forum der Letteren 2*, 110–125.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673–2681.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, *96*(4), 523.

Skousen, R. (1989). *Analogical modeling of language*. Springer Science & Business Media.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Vaswani, A. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Yang, X., Chen, J., van Eerden, A., Samin, A., & Bisazza, A. (2023, May). Slaapte or sliep? extending neural-network simulations of English past tense learning to Dutch and German. In T. Alumäe & M. Fishel (Eds.), *Proceedings of the 24th nordic conference on computational linguistics (nodalida)* (pp. 92–102). University of Tartu Library. https://aclanthology.org/2023.nodalida-1.11

Appendix A

Hyperparameter setting results from Experiment 1

Network size		Batch	Learning		Train			Dev			
Emb/LSTMs	Dropout	size	rate	All	Irreg	Reg	All	Irreg	Reg	r	ρ
TYPE$_{\text{SGT}}$											
300/100	0.1	32	0.001	99.31	92.02	99.70	95.10	51.82	97.39	.53	.49
300/100	0.3	32	0.001	98.75	81.12	99.75	95.56	43.18	98.29	.56	.51
300/100	0.5	32	0.001	98.69	79.44	99.78	95.80	44.55	98.44	.57	.53
400/200	0.1	32	0.001	98.26	76.97	99.43	95.19	40.91	98.02	.54	.53
400/200	0.3	32	0.001	98.37	78.54	99.43	95.32	43.64	97.99	.52	.48
400/200	0.5	32	0.001	98.01	70.00	99.59	95.66	40.46	98.54	.58	.55
300/100	0.5	16	0.001	98.73	81.91	99.67	95.95	47.73	98.47	.55	.54
300/100	0.5	64	0.001	98.14	68.76	99.78	95.70	40.91	98.59	.56	.52
400/200	0.5	16	0.001	97.73	64.27	99.61	95.78	36.82	98.74	.52	.56
400/200	0.5	64	0.001	98.27	77.08	99.43	95.63	45.00	98.29	.57	.55
400/200	0.5	32	0.0001	98.49	77.42	99.67	88.59	37.72	91.30	.53	.45
400/200	0.5	32	0.01	95.09	37.75	98.30	94.53	35.00	97.61	.59	.62
400/200	0.5	64	0.0001	97.83	66.52	99.57	88.13	32.27	91.08	.38	.41
400/200	**0.5**	**64**	**0.01**	**97.05**	**57.53**	**99.24**	**95.44**	**41.36**	**98.24**	**.61**	**.59**

Average accuracy on real verb data and Pearson's r and Spearman's ρ with human Wug Test data: hyperparameter settings Experiment 1. (Part 1 of 2)

Network size		Batch	Learning		Train			Dev			
Emb/LSTMs	Dropout	size	rate	All	Irreg	Reg	All	Irreg	Reg	r	ρ
TOKEN$_{\text{SGT}}$											
300/100	0.1	32	0.001	99.65	97.42	99.79	95.36	47.72	97.96	.51	.47
300/100	**0.3**	**32**	**0.001**	**99.36**	**93.82**	**99.68**	**95.21**	**48.18**	**97.74**	**.61**	**.53**
300/100	0.5	32	0.001	98.96	91.35	99.39	95.85	49.10	98.39	.53	.47
400/200	0.1	32	0.001	99.08	94.04	99.37	95.10	51.36	97.46	.61	.50
400/200	0.3	32	0.001	99.35	96.07	99.55	95.57	51.82	97.99	.61	.51
400/200	0.5	32	0.001	99.11	91.46	99.54	95.66	45.00	98.39	.45	.42
300/100	0.3	16	0.001	99.03	90.90	99.47	95.19	45.00	97.84	.40	.39
300/100	0.3	64	0.001	99.64	96.52	99.81	95.24	48.18	97.74	.50	.46
400/200	0.3	16	0.001	98.34	85.84	99.05	95.12	48.64	97.51	.50	.46
400/200	0.3	64	0.001	99.19	94.94	99.42	94.71	48.18	97.21	.57	.51
300/100	0.3	32	0.0001	99.91	99.21	99.96	88.61	42.27	91.23	.55	.44
300/100	0.3	32	0.01	95.19	65.06	96.79	93.74	43.64	96.48	.53	.53
400/200	0.3	32	0.0001	99.98	99.78	99.99	86.41	39.55	89.12	.46	.35
400/200	0.3	32	0.01	96.38	75.06	97.59	93.35	45.91	95.95	.50	.51
TYPE$_{\text{MTSK}}$											
300/100	0.1	32	0.001	97.57	61.23	99.63	94.85	38.18	97.84	.57	.53
300/100	0.3	32	0.001	96.76	45.96	99.62	95.41	37.27	98.42	.62	.60
300/100	0.5	32	0.001	97.80	66.63	99.53	95.99	42.27	98.72	.56	.54
400/200	0.1	32	0.001	97.69	69.33	99.26	94.68	40.91	97.53	.50	.56
400/200	0.3	32	0.001	97.22	59.89	99.30	95.63	42.27	98.37	.48	.52
400/200	0.5	32	0.001	97.70	66.85	99.43	95.75	41.36	98.54	.51	.47
300/100	0.1	16	0.001	97.47	58.88	99.65	95.55	40.91	98.39	.51	.53
300/100	0.1	64	0.001	97.50	60.45	99.59	93.20	35.00	96.26	.49	.43
300/100	0.3	16	0.001	98.10	73.71	99.48	94.88	46.36	97.41	.50	.49
300/100	0.3	64	0.001	97.54	63.93	99.44	93.33	42.73	96.08	.44	.42
300/100	0.1	32	0.0001	90.48	28.76	94.07	79.24	27.27	82.19	.27	.28
300/100	0.1	32	0.01	97.05	62.92	98.98	94.68	40.91	97.54	.46	.48
300/100	0.3	32	0.0001	90.01	26.85	93.74	81.02	27.73	83.94	.34	.25
300/100	**0.3**	**32**	**0.01**	**97.13**	**53.26**	**99.58**	**95.32**	**36.82**	**98.27**	**.66**	**.61**
TOKEN$_{\text{MTSK}}$											
300/100	0.1	32	0.001	99.37	94.38	99.66	94.27	43.64	97.96	.50	.48
300/100	0.3	32	0.001	99.12	93.25	99.43	94.64	45.91	97.24	.54	.47
300/100	0.5	32	0.001	98.68	88.43	99.25	94.56	44.55	97.21	.53	.48
400/200	0.1	32	0.001	98.67	90.90	99.11	94.27	42.18	97.06	.40	.38
400/200	0.3	32	0.001	99.03	92.70	99.40	95.17	45.45	97.86	.44	.42
400/200	0.5	32	0.001	98.46	86.52	99.08	94.83	44.55	97.49	.52	.47
300/100	0.3	16	0.001	99.58	98.09	99.67	94.51	47.27	97.01	.46	.47
300/100	0.3	64	0.001	98.75	91.01	99.21	94.81	50.45	97.24	.48	.43
300/100	**0.5**	**16**	**0.001**	**99.0**	**92.25**	**99.39**	**95.39**	**50.00**	**97.86**	**.58**	**.48**
300/100	0.5	64	0.001	98.96	88.76	99.54	94.83	43.63	97.61	.37	.44
300/100	0.3	32	0.0001	99.05	93.37	99.38	88.52	44.09	91.06	.52	.43
300/100	0.3	32	0.01	96.91	76.07	98.09	93.81	44.09	96.51	.54	.47
300/100	0.5	16	0.0001	97.92	93.93	98.18	89.20	43.64	91.76	.54	.44
300/100	0.5	16	0.01	94.28	58.54	96.23	92.35	42.27	95.00	.38	.41

Average accuracy on real verb data and Pearson's r and Spearman's ρ with human Wug Test data: hyperparameter settings Experiment 1. (Part 2 of 2)