Compositional Morphology Through Deep Learning

A thesis presented

by

Ekaterina Vylomova

ORCID: 0000-0002-4058-5459

to

School of Computing and Information Systems

in total fulfillment of the requirements

for the degree of

Doctor of Philosophy

The University of Melbourne

Melbourne, Australia

November, 2018

Supervised by:

Tim Baldwin and Trevor Cohn

Declaration

This is to certify that

(i) the thesis comprises only my original work towards the PhD except where indicated in the Preface,

(ii) due acknowledgement has been made in the text to all other material used,

(iii) the thesis is less than 100, 000 words in length, exclusive of tables, maps, bibliographies and appendicies.

Ekaterina Vylomova

November, 2018

Abstract

Most human languages have sophisticated morphological systems. In order to build successful models of language processing, we need to focus on morphology, the internal structure of words. In this thesis, we study two morphological processes: inflection (word *change* rules, e.g. `run`→*`runs`*) and derivation (word *formation* rules, e.g. `run`→`runner`).

We first evaluate the ability of contemporary models that are trained using the distributional hypothesis, which states that a word's meaning can be expressed by the context in which it appears, to capture these types of morphology. Our study reveals that inflections are predicted at high accuracy whereas derivations are more challenging due to irregularity of meaning change. We then demonstrate that supplying the model with character-level information improves predictions and makes usage of language resources more efficient, especially in morphologically rich languages.

We then address the question of to what extent and which information about word properties (such as gender, case, number) can be predicted entirely from a word's sentential content. To this end, we introduce a novel task of contextual inflection prediction. Our experiments on prediction of morphological features and a corresponding word form from sentential context show that the task is challenging, and as morphological complexity increases, performance significantly drops. We found that some morphological categories (e.g., verbal tense) are inherent and typically cannot be predicted from context while others (e.g., adjective number and gender) are contextual and inferred from agreement. Compared to morphological inflection tasks, where morphological features are explicitly provided, and the system has to predict only the form, accuracy on this task is much lower.

Finally, we turn to word formation, derivation. Experiments with derivations show that they are less regular and systematic. We study how much a sentential context is indicative of a meaning change type. Our results suggest that even though inflections are more productive and regular than derivations, the latter also present cases of high regularity of meaning and form change, but often require extra information such as etymology, word frequency and more fine-grained annotation in order to be predicted at high accuracy.

To my loving parents, Galina and Aleksei.

Acknowledgements

First and foremost, my highest gratitude goes to my great supervisors, Tim Baldwin and Trevor Cohn, my "Cyril and Methodius", who introduced me to a whole new world of academic life and always supported me in all my endeavours. During my PhD course, I was fortunate to meet and collaborate with many great researchers. A large part of my research was done together with my good friends, Ryan Cotterell and Christo Kirov, who always inspired me to learn and explore more. I am also grateful to people with whom I collaborated during the JSALT'15 workshop, Reza Haffari, Kaisheng Yao, Chris Dyer and Kevin Duh, from whom I learned a lot about neural machine translation. I am also extremely thankful to Jason Eisner for all his support and brainstorming. Many thanks go to Laura Rimell with whom we worked on my first paper, which was eventually presented at ACL 2016. Certainly, this work would not be complete without David Yarowsky and his passion for derivational paradigms, and I hope we will have a chance to continue research in this direction. I was also very pleased to visit Hinrich Schütze and Alex Fraser's labs in Munich and Chris Biemann's group in Darmstadt in 2016, both of which great experiences, and I am very thankful to them for making this happen.

Although my MSc and BSc are both in Computer Science, I have always been fascinated by linguistics. Here, I would like to thank all linguists with whom I had a chance to meet and be involved in fruitful discussions during the CoEDL Summer School 2017: Nick Evans, Mark Ellison, Martin Haspelmath, Sabine Stoll, Balthasar Bickel, and John Mansfield.

During the last three years, I was also involved in teaching various university subjects, and would like to thank people with whom we tried our best to do "knowledge transfer" to

new generations of students: Jeremy Nicholson, David Eccles, Greg Wadley, Ben Rubinstein, Sue Wright, Rao Kotagiri, and Karin Verspoor.

I would like to thank our great NLP lab and people with whom I shared a significant part of my PhD path: Bahar Salehi, Long Duong, Afshin Rahimi (and his wife Ava), Oliver Adams, Daniel Beck, Matthias Petri, Fei Liu, Doris Hoogeveen, Nitika Mathur, Mohammad Oloomi, Xuanli He, Brian Hur (and his wife Lena), Yitong Li, Shiva Subramanian, Ned Letcher, Moe Fang, Vu Hoang, Yuan Li, Malcolm Karutz, Anirudh Joshi, Julian Brooke, Philip Schulz, Alex Vedernikov, Michał Lukasik, Simon De Deyne, Luis Soto, Sasha Panchenko, and many-many others.

And, certainly, this thesis would not have been be possible without the support of my family, especially Andrei and Mark, who always believed in my abilities.[1]

Finally, I would like to thank my awesome committee chair, Alistair Moffat, and (of course) Google Australia for supporting my thesis research.

[1] and even sometimes overestimated them (:

Publications

Part of the material presented in the thesis has been published before. In particular, Chapter 3 is based on publications [1,2], and most of the material for Chapter 5 was derived from [3].

[1] Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1671–1682, 2016.

[2] Ekaterina Vylomova, Trevor Cohn, Xuanli He, and Gholamreza Haffari. Word representation models for morphologically rich languages in neural machine translation. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 103–108, 2017a.

[3] Ekaterina Vylomova, Ryan Cotterell, Timothy Baldwin, and Trevor Cohn. Context-aware prediction of derivational word-forms. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: (Volume 2 : Short Papers)*, pages 118–124, 2017b.

Contents

List of Figures

List of Tables

Abbreviations

1SG/PL	1st person (singular/plural)
2SG/PL	2nd person (singular/plural)
3SG/PL	3rd person (singular/plural)
ACC	accusative
ADJ	adjective
ADV	adverbial
AGEN	agentive
AN	animate
ASP	aspect
AUX	auxiliary
DAT	dative
DEF	definite
DET	determiner
DU	dual
ESS	essive
F	feminine
FUT	future tense
GEN	genitive
GER	gerund
IMP	imperative
IMPF	imperfective
INAN	inanimate
IND	indicative
INDEF	indefinite
INS	instrumental
IPFV	imperfective
LOC	locative
M	masculine
N	neuter
NOM	nominative
NP	noun phrase
PART	partitive
PFV	perfective
PREP	prepositional
PRES	present tense
PL	plural
PTCP	participle
SG	singular
V	verb
VOC	vocative
VP	verbal phrase
*	ungrammatical or proto-form

Chapter 1

Introduction

1.1 Context of the Study

Consider the following Russian sentence:[1]

(1) Glok*aja* kouzdr*a* štek*o* budlan*ula*
Glocky.F.SG.NOM kouzdress.F.SG.NOM steckly brut.PAST.3SG
bokr*a* i kurdjač*it* bokrënk*a*.
bock.M.SG.ACC and cudder.PRES.3SG bockling.M.SG.ACC
"The glocky kouzdress steckly brutted the bock and is cuddering the bockling."

Although in this nonsense sentence the words have nonce stems, they have semantically plausible suffixes. Even though we have never seen them before, the English grammar and, in particular, our awareness of syntax and morphology of English allows us to reconstruct its possible meaning. For instance, we may conclude that some creature, kouzdress, is the subject of the sentence, and it has a characteristic of being glocky. The kouzdress has done something in the past to another creature, the bock, and now is continuously *cuddering* the bock's offspring, the bockling.

In English, it is not evident that the bock is an animate object, but in the Russian version, it

[1] Example introduced in Russian by Lev Shcherba (Uspensky, 1956) and translated here. Suffixes are *underlined*.

is clear from a specific declension form that signifies that it is a single animate creature. In addition, Russian morphology also marks gender on nouns, adjectives, and verbs. Therefore, in the case of Russian, we will also infer the gender of all participants.

Now let's turn to language production. Suppose there is a hypothetical English verb `to compsognate`. For an English-speaker it is straightforward to conjugate the word into past, future, or present tense. Most likely the corresponding forms will look like *`compsognated`*, *`will compsognate`*, and *`compsognate`*. The speaker may even form a new lexeme, `compsognatable`, to express the meaning of "ability to be compsognated", and transform it into a noun such as `compsognatability`. The speaker might be less confident about the words to signify "someone who compsognates" which might be `compsognater`, `compsognator` or even `compsognatant`.

The ability to generate new word forms and infer the meanings of derived forms is essential, especially when dealing with morphologically rich and resource-poor languages,[2] where most words will not have explicit labels. Therefore, in this thesis, we aim to take a closer look at morphology, the study of the anatomy of words, by exploring various models of morphology and how they can be used to improve general natural language processing ("NLP"). One of the main research questions we are trying to answer in this thesis is how to build a model that will be able to learn the inherent lexical structure and semantics implicitly from a string input.

Most successful contemporary NLP models are so-called "deep learning models", which is a class of neural models with vast parameter space and advanced training algorithms. They have achieved breakthrough performance improvements in various areas, such as image and speech recognition (Hinton et al., 2012; Krizhevsky et al., 2012), language modelling (Bengio et al., 2003; Mikolov et al., 2010; Mnih and Teh, 2012), and machine translation (Bahdanau et al., 2015; Cho et al., 2014; Sutskever et al., 2014). In terms of NLP, one of the downsides of these models is their application to words rather than sub-word units. Some

[2]We consider language to be "high-resource" if it has a large amount of annotated data such as parallel and tagged corpora, e.g. as we have for English, German, or French. Many languages are less documented and may only have a very limited amount of monolingual data or a small number of documented dictionary entries. Such languages are referred to as "low-resource".

models such as *word2vec* (Mikolov et al., 2013a) and *GloVE* (Pennington et al., 2014) are based on **distributional semantics** approach that goes back to Firth's assumption, "you shall know a word by the company it keeps" (Firth, 1957, p. 11). Although the idea of representing a word's meaning as a set of its possible contexts is quite useful, still the representation could be enriched with information about the composition of the word itself. For instance, awareness that `reproducibility` should be decomposed into `[[re- + produce] + -able] + -ity` that first takes the verb `to produce` then attaches to it a prefix `re-` which, in this case, states that the action is repeated. Then the "ability" suffix `-able` is added, and, finally, the suffix `-ity` transforms the adjective to its nominal form. The idea of distributional semantics can be also applied to the inner structure of a word. This supports better modelling of morpheme meanings and functions as well as enables the learning of their productivity and regularity. For instance, potentially the model might also infer that `-able` suffix attaches to verbs forming adjectives with "ability" meaning.

Morphology, the linguistic study of the internal structure of words, has two main goals: (1) to describe the relationship between different words in the lexicon; and (2) to decompose words into *morphemes*, the smallest linguistic units bearing meaning. Here we can identify two key processes, **inflection** and **derivation**, corresponding to word change and word formation, respectively.

The first one, **inflection**, primarily marks features that are necessary for syntax, e.g. case, gender, tense, number. For example, the Russian form *pribyvšemu* (English "to the arrived") corresponds to a combination of features such as past participle, perfective aspect, singular number, masculine/neuter gender, and the dative case. In most languages inflection does not change the part of speech of the word[3] and tends not to change its basic meaning. The set of inflectional forms for a given lexeme is said to form a paradigm, e.g., the full paradigm for the English verb `to take` is ⟨*take*, *taking*, *takes*, *took*, *taken*⟩. Each entry in an inflectional paradigm is termed a slot and is indexed by a syntactic-semantic category, e.g., the PAST form of *take* is *took*. We may reasonably expect that all English

[3]There are some cases in Australian Aboriginal languages (Evans, 1995) where POS changes (probably due to their polysynthetism).

Case	Singular	Plural
NOMINATIVE	*książka*	*książki*
GENITIVE	*książki*	*książek*
DATIVE	*książce*	*książkom*
ACCUSATIVE	*książkę*	*książki*
INSTRUMENTAL	*książką*	*książkami*
LOCATIVE	*książce*	*książkach*
VOCATIVE	*książko*	*książki*

Table 1.1 An example of inflectional paradigm. Declension of Polish word `książka` ("book").

verbs—including neologisms—have these five forms.[4] Furthermore, there is typically a fairly regular relationship between a paradigm slot and its form (e.g., add `-s` for the third person singular form to the English verb in the present tense as in *thinks*, *produces*, *writes*).

The fact that English presents quite a limited inflectional system helps to explain why morphology has received less attention in the computational linguistics and natural language processing literature than it is arguably due, given the English-centrality of the field. Table 1.1 exemplifies a moderate-sized inflectional paradigm table for Polish noun declension for comparison. Most languages of the world present rich morphological systems. The linguistic typology database WALS shows that 80% of the world's languages mark verb tense through morphology while 65% mark grammatical case (Haspelmath et al., 2005).

To summarise, inflections are highly regular, productive, paradigmatic and mandatory, i.e. they apply to every stem and have to be expressed.

The second process, **derivation** is one of the key processes by which new lemmata are created. For example, the English verb `corrode` can be expressed as the noun `corrosion`, the adjective `corrodent` or `corrosive`, and numerous other complex derived forms such as `anticorrosive`, `corrosiveness`, `corrosively`, `corrosivity`. This contrasts with inflectional morphology which produces grammatical variants of the same core lexical item (e.g., *take* $\mapsto$ *takes*). Derivational morphology is often highly productive, lead-

[4] Only a handful of irregular English verbs distinguish between the past tense and the past participle, e.g., *took* and *taken*, and thus have five *unique* forms in their verbal paradigms; most English verbs have four *unique* forms. English copulas form a standalone paradigm that consists of *am*, *is*, *are*, *were*, *was*.

ing to the ready creation of neologisms such as Kroneckerise and Kroneckerisation, both originating from the Kronecker product.

Derivations present less regularity and therefore are not usually seen as paradigmatic. Some of the derived words start being perceived as a complete meaning-bearing unit over time and hence might be non-compositional, which makes derivation modelling an extremely challenging topic. For instance, both growth and warmth originated at least from Middle English and were initially formed from verbs and adjectives by attaching the −th suffix (or its proto-versions ⋆−þu,⋆−þ)[5] which indicates abstract meaning. Although words with −th are actively used in contemporary English, all of them are perceived as a single unit and −th is seen as a part of the stem. The suffix itself has ultimately become non-productive, i.e. is not used for neologism creation. Hence we refer to calmness and kindness, but not calmth and kindth. Table 1.2 presents a partial attempt at a paradigmatic table for English verbal nominalisations.[6] The table illustrates irregularity in meaning and form: some slots are unfilled for a given lexeme. This might be explained by restrictions on the initial lexeme as in the case of "Patient", or a blocking effect, i.e. a situation when for the same concept there exists another lexeme already (e.g., marrier might be blocked by fiancé, or in Russian svinënok ("piglet" originating from ⋆svīnī) is blocked by porosënok ("piglet" originating from ⋆porsę)). As for the form, we might also notice a greater variation in some slots (e.g., RESULT can be realised using different suffixes such as −ion, −ation, −ment, −ence, etc.). Unlike in inflections, there are also often more alternative, or competing forms, such as nominals move and movement. Note that the second lexeme additionally expresses various specific meanings. Therefore, this makes derivations more challenging to formalise than inflections.

To summarise, in this thesis we aim to create a model that will be suitable for predicting inflectional as well as derivational forms from a word's constituent parts, i.e. generating their forms from their meaning description, e.g. a paradigm slot or a sentential context corresponding to it. To do so, we utilise several contemporary neural architectures. The

[5] Based on Douglas Harper, Online Etymology Dictionary, 2001–2018

[6] The partial paradigm structure was motivated by linguistic studies and proposed in Cotterell et al. (2017b).

Verb Base/Suffix	-er/-or	-ee	-ment/-tion	-ive/-ent	-able/-ible
POS	V↦NOUN	V↦NOUN	V↦NOUN	V↦ADJ	V↦ADJ
Semantic	AGENT	PATIENT	RESULT	CHARACTERISTIC	POTENTIAL
animate	animator	--	animation	--	animatable
advise	adviser	advisee	advice	--	advisable
educate	educator	educatee	education	educative	educable
teach	teacher	--	teaching	--	teachable
amputate	--	amputatee	amputation	--	--
attract	attractor	attractee	attraction	attractive	attractable
--	aggressor	aggressee	aggression	aggressive	--
employ	employer	employee	employment	--	employable
move	mover	--	movement	--	movable
place	placer	--	placement	--	placeable
escape	escapee	--	--	--	escapable
corrode	corroder	--	corrosion	corrosive	corrosible
derive	deriver	derivee	derivation	derivative	derivable
marry	--	--	marriage	--	marriagable
eat	eater	--	--	--	edible
codify	codifier	--	codification	--	codifiable
think	thinker	--	thought	--	thinkable
prohibit	prohibitor	prohibitee	prohibition	prohibitive	prohibitable
cook	cook	cooker	cooking	--	cookable
eat	eater	--	eating	--	edible

Table 1.2 Possible partial derivational paradigm for several English verbs; semantic gaps are indicated with −−. Note that suffixes often significantly vary within a single slot.

proposed model should be a well-suited solution for the large lexicon problem that often occurs in various tasks in morphologically rich languages.

1.2 Aim and Scope

Nearly every traditional NLP task (for example, machine translation, language modelling, part of speech tagging) has to deal with out-of-vocabulary words, i.e. the words or forms that the system did not observe or observed only a few times during training stage. A large percentage of them are just morphological variations of known stems or related word forms. This problem is highly important for synthetic languages, i.e. those presenting rich morphology. Figure 1.3 illustrates the number of forms of the paradigm for `čéška` observed given different sizes of corpus. For instance, the nominative case, as soon as it signifies a subject of the sentence, is observed in all corpora sizes. The accusative case (marking an object) is less likely to be present, but its form in this particular case matches the form of the

Case	Surface Form	50K	500K	5M	50M
NOMINATIVE	*čéšky*	•	•	•	•
GENITIVE	*čéšek*	–	•	•	•
DATIVE	*čéškám*	–	–	•	•
ACCUSATIVE	*čéšky*	○	○	•	•
VOCATIVE	*čéšky*	○	○	○	○
LOCATIVE	*čéškách*	–	•	•	•
INSTRUMENTAL	*čéškami*	–	–	–	•

Table 1.3 Morphological variants of the Czech lemma `čéška` for different corpus sizes. Here • indicates that the variant occurs, and ○ – that the same surface form appears but it corresponds to another morphological feature combination (Huck et al., 2017).

nominative case, and, therefore, we observe it even in the smallest corpus. The instrumental case is much less likely to be observed for every lemma, and for this reason models will have to generalise in order to understand this novel form by seeing other lemmata that follow the same paradigm.

It is also important to underline here that we focus on compositional cases, i.e. cases where the meaning of the whole can be predicted or formed from the meaning of its parts. Inflectional morphology is highly compositional by its nature. Derivations, on the other hand, are more complex. As mentioned earlier, some derivational morphemes become non-productive and the words themselves get lexicalised and typically processed as a single unit by native speakers (e.g., `action`, `actor`, `actual`, `active` all originating from the Latin "do, perform", or `thriller` referring to "a thrilling, suspenseful book or movie"). Therefore, their meanings are often completely different from what could be inferred from their parts or might be associated with some semantic restriction that cannot be predicted from a simple compositional model. In this way, they are close to multiword expressions (Sag et al., 2002), which is out of scope of this work.

Bearing this in mind, for inflected and derived forms we aim to estimate a representation given either their tags and lemma or base forms, respectively, or some contextual representation corresponding to a particular tag combination. The main question here is how well we can perform such prediction and which compositionality function better combines the parts and predicts syntactic and semantic features. In order to do so, we investigate various

contemporary neural models. The task is decomposed into the following constituent parts, framed by individual research questions:

RQ1: What information do models trained based on the distributional semantics hypothesis capture?

Most contemporary language models' objective functions are based on prediction of words from their (often sentential) contexts. Here we investigate what types of relations between words are captured by this training strategy. In particular, we look at various types of binary relations such as hypernymy (`lion` → `animal`) for lexical semantics, nominal plurality (*`lion`* → *`lions`*) as an example of morphosyntax, and repetition (`discover` → `rediscover`) for morphosemantics. There are many ways to express such relations using vector space models, and we focus on the vector difference (**lions** − **lion**) approach, which has been shown to work well elsewhere (Mikolov et al., 2013c).

RQ2: Do character-level models provide better representations of morphological similarity than word-based? Which neural architecture better expresses morphological information?

Human annotated features such as morphological boundaries are extremely valuable, but cases when they are available are quite rare and usually limit the scope of languages to well-documented cases only. Therefore, one possible solution to this problem is to consider character-level representations. Here, we investigate how well various neural architectures capture morphological information as well as whether character-level segments could be compared to morphemes. We compare two popular contemporary neural models, recurrent (Elman, 1990) and convolutional (LeCun and Bengio, 1995) and evaluate their morphology awareness across several typologically diverse languages.

RQ3: How well can derived and inflected forms be predicted directly from a sentential context?

Earlier studies of inflectional morphology presented in the shared task on morpholog-

ical reinflection across over 50 languages (Cotterell et al., 2017a) showed that there is a lot of regularity in inflectional processes, even in languages with highly complex morphology; and the neural systems outperform non-neural ones and achieve high accuracy results on most languages in data greedy conditions. In this task, researchers relied on linguistic knowledge of paradigmatic slots. In a more realistic scenario, however, the system has to predict the forms directly from the context. First, we test how well the model is able to predict information about gender, case, and tense, and ask the question: how coherent are the model's predictions? This question has a linguistic basis because little is known about the internal nature of case systems. In addition, we study various languages and compare the models' uncertainty. Do languages with rich morphology such as Hungarian require more capacity and lead to higher entropy?

Second, we test the same idea on derivations. Derivations present semantic changes to the base form. How well can we predict these semantic shifts at the token level based on surrounding context? Here we propose an encoder-decoder model which is trained to capture morphotactics to be able to produce surface forms and learn a mapping from contextual information to suffix semantics. We also compare our results to those obtained on inflections.

Finally, in terms of the thesis, we only look at contemporary state of language and leave a diachronic perspective, i.e. its development over time, for future work.

1.3 Thesis Structure

The thesis is structured as follows. First, **Chapter 2** provides linguistic information on the nature of morphological processes and describes the key concepts that the thesis builds on. Then we continue with a brief summary of traditional models of morphology proposed within theoretical linguistics and computer science. We start with inflectional morphology and present a description of various existing approaches to it. Then we move to derivational morphology providing a detailed analysis of the relations which are typically expressed by means of derivations. Then the chapter progresses to contemporary ideas of paradigmatic

treatments of derivation, and theories that place both inflections and derivations on a single continuous scale. At the end of the first part of the chapter we outline a number of tasks existing in morphology modelling. The second part of the chapter discusses contemporary approaches to distributional semantics. It provides a number of vector space models proposed during the last two decades and lists a set of tasks which allow us to run a comparison of the models. The final part of the chapter introduces the notion of compositionality and how contemporary neural models such as recurrent (Elman, 1990), recursive (Socher et al., 2013b), and convolutional (LeCun and Bengio, 1995) networks express compositional functions over characters, morphemes, and words.

Chapter 3 is devoted to the analysis of distributed word representations. The chapter addresses *RQ1* and *RQ2*, i.e. investigates the word-level information captured by distributed representations and compares various levels of representation. In particular, it provides an analysis of word vectors obtained in two tasks: language modelling and machine translation. We compare several word- and character-level models, including the effect of pre-trained and learned end-to-end. Each model is evaluated in terms of how well it represents morphological similarities as well as its ability to differentiate various relations (hypernyms, meronyms, etc.). Our main finding is that morphosyntactic information is typically learned better than morphosemantic and lexical.

Chapter 4 addresses *RQ3*. In particular, it investigates modelling of inflectional paradigms and focuses on contextual inflection prediction, a less data intensive and more challenging setting where the morphological tags are not explicitly provided, but rather have to be inferred from a sentential context. We propose several models, namely an encoder-decoder model and neural conditional random fields, and provide an analysis of each model's performance. We compare different languages in terms of their morphological complexity and language model prediction accuracy. We also show that some grammar categories, such as verbal gender and number, are contextual whereas others like verbal tense are inherent, and lead to higher uncertainty. We additionally evaluate the models' accuracy on agreement prediction. The end of the chapter compares the setting with the contextual inflection subtask of the SIGMORPHON 2018 shared task and discusses the results of this task.

Chapter 5 continues with *RQ3* and discusses derivational morphology. The first part of the chapter is devoted to derivational paradigms. We first apply the task of inflectional paradigm completion to derivations that aims at addressing one of the main research questions, i.e. whether derivations can be viewed as paradigmatic. The results obtained with an encoder-decoder model show that regular and productive derivational transformations can be predicted with high accuracy, although less regular and productive transformations are still very challenging. The second part of the chapter addresses the contextual prediction of derivations. There, we try to generate derived forms from their sentential context and base form and observe the same pattern as before, namely that regularity and productivity plays a significant role. In order to address irregularities, we propose several treatments such as the inclusion of etymological and frequency information to further improve the models.

Finally, we summarise the results and contributions of the previous chapters in **Chapter 6**, and outline possible future directions of research.

Chapter 2

Background

This chapter provides necessary background on language and, in particular, morphology. We describe two types of morphology, inflectional and derivational, and how both of them are addressed in linguistics and computational linguistics. We then provide a background on the word- and character-level neural models we will further evaluate.

2.1 Language

Human language has always been at the centre of philosophical studies. A great volume of philosophical works from Ancient Greeks and Romans to Wittgenstein had been devoted to various aspects of language such as where the meaning comes from, how a word form relates to its meaning, and how connections and dependencies between words within a sentence are expressed.

What is special about Language? Surprisingly, the term *Language* does not have a generally accepted definition. Typically, linguists refer to it as a communication system. Therefore, in its very broad sense, in addition to humans, language also applies to animals and plants. But in this thesis we will focus on human language, hence we first need to define it. So, what is special about human language? According to Hockett (Hockett, 1958, 1977; Hockett and Altmann, 1968; Hockett and Hockett, 1960) in the context of a comparative

analysis of animal and human language, there are approximately 10 essential properties unique to humans such as semanticity, productivity, displacement, recursion, discreteness, duality of patterning, hierarchy, reflexiveness.

We take a closer at those relevant to *morphology*, the main focus of the current thesis.

First of all, **semanticity** is the property whereby some elements of language refer to objects of the real world. For instance, `potato` refers to a particular type of vegetable. At the same time, there are such elements that stand for a whole class of similar objects, such as `flower` that signifies any flower. Some elements, such as morphemes, might not stand for any real objects, but rather generalise to more abstract concepts. For instance, `-s` in *cats* is a sign of plurality. As Burlak (2017) notes, it is essential that semanticity of communication system requires the signals to be detached from the objects they refer to, and therefore it is closely related to arbitrariness of sign.

Another important feature, **productivity**, refers to the ability to express and produce an unlimited number of new utterances from a limited number of initial elements. Clearly, children do not only memorise utterances but are able to generate new ones based on rules and patterns they learn from adult speech. U-shaped language development (Ervin, 1964) and Nicaraguan sign language (Senghas et al., 2004) emergence are good examples of human productivity. In particular, in the first case children attempt to generalise over patterns they see in the language and reconstruct the grammar, leading to over-generalisation and production of regular forms in place of irregular ones (for instance, *runned* instead of *ran*). Nicaraguan sign language evolved as a by-product of interaction between deaf children. The children initially knew only gestures that they used at home (i.e. the gestures were individual-specific and idiosyncratic), and after they all started school, the gestures became more unified and year-by-year structure appeared.

Hauser et al. (2002) state that **recursion** is the main and unique component of human language.[1] The mechanism of **recursion** enables any sentence to be easily embedded into a larger phrase, for instance, starting with `I thought that ...`. This holds for morphology as well, i.e. a stem could possibly serve as a base for a new word. For instance,

[1]This statement significantly differs from Chomsky's earlier position with respect to language expressed in Chomsky (1975).

`read`→ `re-read` → `re-re-read`. This could be easily extended further if there is a pragmatical necessity to express such a meaning. Similarly, in some languages, we can form a potentially infinite chain of diminutive derivations. For instance, in Slovak it is possible to use suffix reduplication to form an adjective `malilililinký` meaning "very-very-very-very small" (Körtvélyessy, 2014).[2] Recursion presents in most, if not all, human languages. There is an ongoing debate about Pirahã, a tribal language of the Amazon River region. One of the main linguists studying this language, Everett, states (Everett, 2009; Everett et al., 2005) that the people speaking this language do not use recursion, and this poses a serious question for the statements mentioned above.

Pinker and Jackendoff (2005) presented a critical view of the ideas expressed in Hauser et al. (2002). If we look at conceptual structure, many concepts are unique to humans only. For instance, the notion of a *week* is entirely based on an ability to count and number abstraction, and it is likely not possible to be learned without language. They also point out that individual elements of language are not organised chaotically but rather form a system. Moreover, the system presents a certain level of **hierarchies**, one for morphosyntax (starting from morpheme → grammatical word → ... → sentence → text) and another one for phonetics (phoneme → syllable → phonetic word → ... → phonetic sentence).[3] Words also relate to each other as hypernyms, synonyms, meronyms and other types of lexical semantic relations on the one hand and morphologically on the other. For instance, we can take all word forms for `run` such as *`run`* – *`runs`* – *`ran`* – *`running`*, or its derivations `runable` – `runner` – `run away` – `run out`, or, alternatively, we can take all the words ending with `-er` and expressing agentive meaning such as `runner` – `jumper` – `mover`. Crucially, Pinker and Jackendoff also note that words' meanings embed information about their compatibility, e.g. as valence in verbs. Some verbs, such as `run` in the meaning of "moving fast by foot", require a single argument (a subject) whereas `put` requires three arguments corresponding to "who?", "what?", and "where?" questions.

[2] English lengthenings such as `cooooooool` are not relevant to recursion, as there is no notion of a discrete morphological process being applied recursively so much as reflection of emphasis-based phonetics in the lexical form.

[3] "A → B" stands for "B is composed of A", i.e. hierarchically A is lower than B.

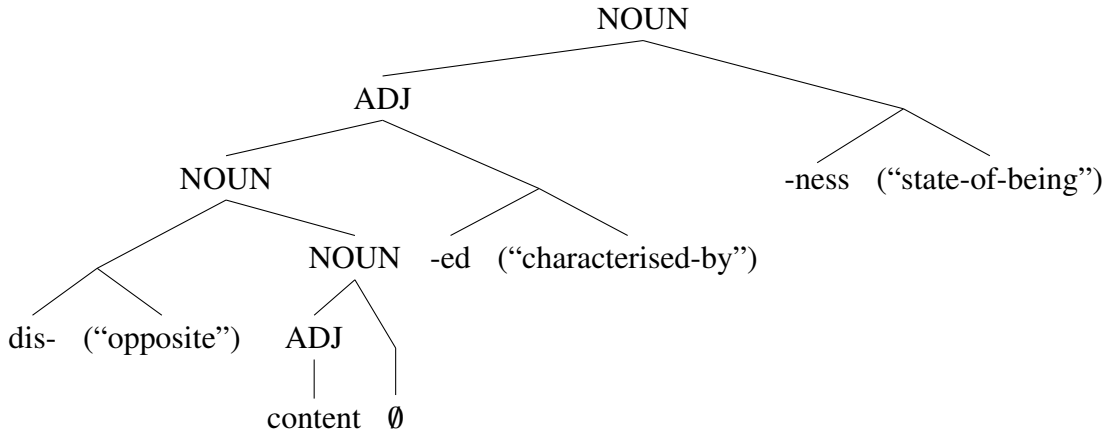
Therefore, a successful model should include all the aforementioned components, i.e. be recursive and the concepts there expressed by words need to form a complex associative network. In the following section, we provide a detailed description of different types of word formation, their interplay, as well as existing theoretical approaches to model the underlying processes.

2.2 Morphology

2.2.1 Linguistic approaches

In the traditional Saussurean view, a linguistic sign consists of two parts: a signified concept (which includes sensory information of visual, auditory, or tactile kind) and a signifier, an acoustic image for it. For instance, `/kæt/` corresponds to a notion of "cat". This unity of the signifier and the signified is a fundamental element in Saussure's semiotic theory. For Saussure, it is uncontroversial that a mapping between them is arbitrary. Hence, a feline creature is referred to in English as `cat`, in French as `chat`, and `neko` in Japanese. Clearly, this statement does not hold true for morphologically complex words with transparent meanings. On the one hand, the signifier is not motivated by the nature of the signified concept, and, therefore, is arbitrary with respect to the concept. But, on the other hand, language is a social phenomenon, so the sign is not purely subjective but rather appears as a result of a consensus in the process of communication. Anderson (1992) exemplifies the relationship between the form and the meaning with `discontentedness` as a "state of being discontented ". `Discontented` is related to "characterised by notable discontent", which could be decomposed into `dis+content`, i.e. opposite to `content`. Finally, the noun `content` could be based on the adjective `content` ("satisfied"). Therefore, he notes, "there are several layers of reference to meaning, rather than a single homogeneous association of a spoken form with some semantic content" (Anderson, 1992, p. 10). Each time a morphologically complex word is decomposed into its constituent parts, and, therefore, there should be a systematic and regular mapping between subparts of the form and corresponding

subparts of the meaning. Anderson provides a good example in order to illustrate the decomposition of the form and the meaning:



The example provides evidence that `discontentedness` is not a single sign, but rather should be considered as a structured composition of some individual simple signs. Structuralists, therefore, proposed a simple model that stated that a) every word is composed of morphemes; b) every morpheme in a form is represented by exactly one allomorph, a variant form of a morpheme; and c) morphemes are organised into hierarchical structures. Further, the generativists' view of word structure was based on these principles. They took the notion of morpheme unchanged; but the generative approach did not specifically address allomorphy (i.e. cases when a single morpheme has multiple surface realisations) instead reducing it to phonological modelling. Similarly, they assumed that principles underlying word-internal structure and morpheme combination could be accommodated by syntax.

A substantial part of recent morphological research has been focusing on the **nature of the form – meaning mapping**: some morphologists proposed it is rule-based similar to phonological rules, while others, as noted above, suggested the use of hierarchical structures analogous to syntax. American structuralists referred them as **Item and Process (IP)** (Boas et al., 1947; Hockett, 1954; Steele, 1995) and **Item and Arrangement (IA)** (as in Bloomfieldian (Bloomfield, 1933), Post-Bloomfieldian (Harris, 1942; Hockett, 1947, 1954), and Generative (Lieber, 1992)) theories, respectively. In the latter case, complex words are decomposed into morphemes, i.e. minimal meaning-bearing units. Therefore, words are only seen as a composition of constituent morphemes (such as `/kæts/` = `/kæt/` + `/s/`).

Item and Arrangement		Item and Process
/kæt/	/s/	/kæt/ → /kæts/
[root]	[plural]	[+N +plural]

Table 2.1 Examples of /kæts/ representation in "Item and Process" and "Item and Arrangement" theories.

This view faces difficulties in the case of irregular forms that could not be analysed in such a way (e.g. `/wənt/` cannot be decomposed and represented as `/go/` + `/ed/`) and have to be recognised as a whole. The former one focuses on a lexical base and processes that add, remove, or modify properties simultaneously affecting the form of the base. Hence, the approach does not suffer from the problem of vowel alternations or irregular forms mentioned earlier.[4] Anderson (1992) and Stump (2001) note that these models are also suitable in order to address derivational morphology, i.e. principles of new word formation.

So far, we briefly highlighted the idea that words can be further decomposed into more basic units, and we also described two general linguistic approaches to word formation, arrangement-based and process-based. In the following section, we provide a typological view on morphology, i.e. we classify languages with respect to their morphological expressiveness.

2.2.2 Morphological Typology

Depending on the way the relations between words are expressed within a sentence, linguists identify three main groups of languages, namely **analytic**, **isolating** and **synthetic**. In **analytic** languages, such as Modern English, relations are mainly expressed by prepo-

[4] Stump (2001) classifies morphological theories into two types: lexical and inferential. In inferential systems, a systematic relation between a lexeme's root and all inflected forms in its paradigm is expressed by a set of rules or formulas. In lexical systems, on the other hand, the mapping between inflectional form marking and a set of its corresponding morphosyntactic properties is similar to an association between a lexeme's root and its grammatical and semantic features. Therefore, in lexical theories a morpheme constitutes a unit of structure, whereas in inferential ones it does not. Second, Stump also provides an extra dimension by identifying theories as realisational or incremental. Both of them identify a connection between morphosyntactic properties (such as noun plurality) and their inflectional exponents (such as endings `-s`/`-es`). In a realisational model, "a word's association with a particular set of morphosyntactic properties licenses the introduction of those properties' inflectional exponents", while in an incremental model "words acquire morphosyntactic properties only as a concomitant of acquiring the inflectional exponents of those properties" (Stump, 2001, p. 3). Further, Stump argues that the inferential/realisational model best fits and describes inflectional morphology.

sitions and particles, although some inflectional morphemes still remain from the declension system of Old English (such as pronoun forms *I/me/mine*, *he/him/his*, and *they/them/their*).

(2) Chinese (Isolating)

wǒmen xué le zhè xiē shēngcí.

I.PL.AN learn .PAST this .PL new word.

"We learned these new words."

(3) Russian (Synthetic)

My vyuč*ili* et*i* nov*yje* slov*a*.

We learn.PAST.PL this.ACC.PL new.ACC.PL word.ACC.PL

"We learned these new words."

On the other hand, as we will demonstrate in Chapter 5, these languages often have a comparatively rich derivational system. **Isolating** languages present an extreme case of that and lack any inflectional morphology. Mandarin Chinese and Vietnamese are examples of this type. In **synthetic** languages, words consist of many morphemes, i.e. multiple concepts here are realised within a single word. For instance, as illustrated in (2), the past tense in Chinese is expressed by a certain particle, le, whereas in Russian, since it is a synthetic language, it is realised together with a plural marker in a single form of the corresponding verb. An extreme case here is **polysynthetism**, i.e. the situation of a word being composed of a large number of morphemes and conveying a sentence-level meaning. Here is an example of West Greenlandic taken from Fortescue et al. (2017):

(4) West Greenlandic (Polysynthetic)

Nannu-n-niuti-kkuminar-tu-

Polar.bear-catch-instrument.for.achieving-something.good.for-PART-

rujussu-u-vuq.

big-be-3SG.INDIC

"It (a dog) is good for catching polar bears with."

Two extra classes are traditionally specified within synthetic languages: **agglutinative** and **fusional**. In the first case, morphemes are concatenated without change or modification in their form, whereas in the second type a single morpheme denotes a combination of grammar tags. All Slavic languages, Sanskrit, Latin, Ancient Greek correspond to the fusional type. Turkic and many Uralic languages, Japanese (with postpositions), and Korean are examples of agglutinative languages.
Consider the following example of agglutinative declension for Estonian:

(5) Estonian (Agglutinative)

`ilus-a-te` `raamat-u-te`

`beautiful.PL.GEN` `book.PL.GEN`

"of beautiful books (Genitive Case, Plural)"

Here, *each* morpheme expresses a particular grammar feature (the genitive case or plural number).

An example of fusional declension (Russian) is:

(6) Russian (Fusional)

`krasiv-yh` `knig+0`

`beautiful.PL.GEN` `book.PL.GEN`

"of beautiful books (Genitive Case, Plural)"

Here, a *single* morpheme (e.g., adjectival `-yh`) simultaneously expresses a combination of grammar features (both the genitive case and plural number).

Another example of fusional morphology is presented by Semitic languages:

(7) Hebrew (Templatic)

`sfarim` `yafim`

`book.PL.NOM` `beautiful.PL.NOM`

"beautiful books (Nominative Case, Plural)"

Here, a pattern (template) is applied to a root (which typically consists of 3 consonants, e.g. `s.f.r` in the case of "book") in order to produce a word form (such as `sefer` for "a book", `hasefer` for "the book", `sfarim` for "books", or even `sifriya` for "library").

2.2.3 Morphology-Syntax Interface: an example of grammatical case

The interplay between morphology and syntax can be best illustrated in terms of grammatical case. **Grammatical case** is one of the most important but, at the same time, vaguely defined grammatical categories that is usually expressed by means of morphology. A canonical case stands for a prototype or an expression of syntactic or semantic role on a noun. Malchukov and Spencer (2009) discuss bounds of the definition. A set of "classical" cases such as the *nominative*, the *accusative*, the *dative* to mark an agent, a direct object and a recipient, respectively, might be extended if different languages are considered. For instance, the *vocative* is typically used to refer to an object being addressed and does not express a syntactic or a semantic role, but is a means of referring to the addressee.

(8) Ukrainian (Slavic)

`Hanno, jdy sjudy!`

`Hanna.VOC come.IMP.SG here!`

"Hanna, come here!"

Another example comes from spatial cases. Peoples living in mountain areas might express various spatial relations such as "up the hill" or "down the hill" morphologically (Creissels, 2009). All these categories greatly extend the case definition bounds.

Still, the notion of grammatical case remains elusive since it is typically hard to assign a particular set of roles to each case. Moreover, the precise number of cases in a language can be controversial and significantly vary.[5] For instance, in Russian it ranges from six to eleven

[5]Contemporary views on grammatical case have been influenced by Greeks who were one of the first to describe the case system. Even our view on the nominative case as being primary and other cases falling out of it comes from the Greeks. The Greeks described their case system, *ptôsis*, which looks quite close to what we have now (Malchukov and Spencer, 2009). And currently the main definition of the case stands as follows, "The grammatical category of the name, expressing its syntactic relations to other words of the utterance or

NOMINATIVE	kartṛ	agent
ACCUSATIVE	karman	object
INSTRUMENTAL	karaṇa	instrument
DATIVE	saṃpradāna	destination/recipient
ABLATIVE	apādāna	source
LOCATIVE	adhikaraṇa	locus

Table 2.2 Cases and Pāṇini's kārakas

(Zaliznyak, 1967).[6] A similar concept, *kārakas*, appeared in Pāṇini's Sanskrit grammar from 600BC to 300BC. Kārakas described relations between nouns and their governing verb and looked more like what we would call semantic roles nowadays: agent, object, instrument, destination, and others. Table 2.2 illustrates the mapping between kārakas and modern cases.

The case category is typically seen as being expressed *morphologically*. Therefore, Modern English is considered to be a case-poor language, since the grammatical case is only left marked on pronouns (compared to Old English in which nouns had singular and plural forms corresponding to five grammatical cases[7]). On the other hand, one can always say that grammatical case roles in English are expressed via prepositions and word order. For example, the utterance `I saw him yesterday` is grammatically correct while `*I saw yesterday him` is not. That would be explained by a locality restriction on case assignment, so the cases are assigned to the first nominal. Similar restrictions apply if pronouns are replaced with nouns. This means that nouns also have *syntactic case* marking that refers to syntactic structure. Such an interplay between morphological and syntactic case is closely related to the problem of distinguishing case allomorphs (such as `/s/`, `/iz/`,`/z/` all corresponding to the plural suffix in English nouns) and different cases. Moreover, in many languages, a single grammatical case might be realised in the form of another case for a specific subset of objects. For instance, the Russian accusative is formally identical

to the utterance as a whole, and also any grammeme of this category (the concrete case)" (Uspensky, 2002, p. 299).

[6] Zaliznyak (1967) developed a mathematical definition of grammatical case as a set of classes of equivalence that was initially proposed by Kolmogorov.

[7] According to Quirk and Wrenn (2002) and "King Alfred's Grammar" available at `http://www.csun.edu/~sk36711/WWW/KAG/`

to either the Russian genitive for animate objects of masculine gender or to the Russian nominative for inanimates, with only the exception of nouns ending with `-a`, `-ja` where it has its own unique form.[8] Many of these differences can be explained from a diachronic perspective. Such variations often start from alternations in syntactic constructions and then develop into the direction of non-canonical objects and subjects, and finally end up as purely morphological matter. As Malchukov and Spencer (2009) also mention, the notion of syntactic case should be motivated syntactically in the sense that it helps to generalise across different declension classes.

2.2.4 Inflections

Inflectional morphology is a set of processes through which the word form outwardly displays syntactic information, e.g., verb tense. It follows that an inflectional affix typically neither changes the part-of-speech (POS) nor the semantics of the word. For example, the English verb `to run` takes various forms: *`run`*, *`runs`* and *`ran`*, all of which convey the concept "moving by foot quickly".

2.2.4.1 Inflectional Paradigm

As mentioned in Section 2.1, concepts are not isolated of each other but rather are related by a wide spectrum of relations. Saussure in his works (Saussure, 1959) referred to them as *associative relations*. For instance, the Russian words *`letaju`* "fly.V.PRES.1SG", *`letala`* "fly.V.PAST.F.SG", and *`letajuščij`* "fly.PART.M.1SG" all share the same lemma. At the same time, we can look from a different perspective and observe regularities in *`letal`* "fly.V.PAST.M.SG", *`kričal`* "cry.V.PAST.M.SG", *`mečtal`* "dream.PAST.M.SG", i.e. a common `-al` suffix that specifies the past tense masculine singular form. Alternatively, we can also group the words by common semantics (hyponyms, synonyms, etc.), or any other feature. One of the most typical views on *paradigm* relates to the groupings based on the common lemma (inflections). So, what is an inflectional paradigm? First, as Pounder (2011) notes, we can focus on patterns specific for a particular lexico-syntactic class. Second, we can

[8] Therefore, it is considered as a "weak" case.

alternatively consider various forms of a single lexeme. For example, a typical English verb may have slots in its inflectional paradigm corresponding to its lemma (`write`), past tense (*`wrote`*), past participle (*`written`*), and third-person singular (*`writes`*) forms. These forms are related by a consistent set of transformations such as suffixation. Traditionally, there is no doubt about the paradigmatic nature of inflections since they present a high level of regularity. Even though there are irregularities, generally all paradigmatic variants are legal. Some of the irregularities can be explained if we look at them diachronically. Table 2.3 illustrates how verbal inflection changed over time. In Old English and Proto-Germanic there were two classes of verbs: strong and weak, as the Grimm brothers referred to them in Grimm (1890). In Old English, the strong verbs produced past forms by means of stem vowel alternation (and their participle typically ended with `-en`), whereas the weak ones attached a suffix. The weak verbs comprise regular verbal inflection (`-ed` suffix) in Modern English. The strong ones correspond to irregular inflections. Another important observation can be made with the strong verbs. Although in Proto-Germanic we see a high level of paradigm regularity, it is not the case in Modern English. Some verbs still end with the suffix `-(e)n` (*`stolen`*, *`broken`*) while others (*`won`*, *`come`*) lost this ending. A closer look at their forms reveals that the `-en` disappeared in verbs having a nasal (`n`, `ng`, `m`) preceded by a vowel in the stem. Although in Modern English the irregular verbs seem to be disconnected, Proto-Germanic inflection was quite regular, but phonetic changes over time made the connections less obvious.

Clearly, such irregular changes are less productive than regular ones in Modern English. This is also supported by observations in child language acquisition when at some stage children begin to over-generalise and apply regular patterns to irregular verbs (U-shaped learning described in Section 2.1). As mentioned in Section 1.2, in this thesis we only focus on the synchronic view on language and leave study of diachronic aspects for future work.

Now, if we look at Russian verbal inflection illustrated in Tables 2.4 and 2.5, we observe that the forms conjugate in person, number, tense, and mood. Each individual paradigm also corresponds to a particular aspect, and perfective and imperfective forms have two different paradigms. Importantly, in the perfective paradigm we observe *empty* slots for some forms

Language/Infinitive		Present, 1SG	Past, 3SG	Past Participle
(S3) to win	Modern English	win	won	won
	Old English	winne	wann	(ġe)wunnen
	Proto-Germanic	⋆winnō	⋆wann	⋆wunnanaz
(S4) to come	Modern English	come	came	come
	Old English	cume	cōm	(ge)cumen
	Proto-Germanic	⋆kwemō	⋆kwam	⋆kumanaz
(S3) to find	Modern English	find	found	found
	Old English	finde	fand	(ġe)funden
	Proto-Germanic	⋆finþō	⋆fanþ	⋆fundanaz
(S3) to rise	Modern English	rise	rose	risen
	Old English	rīse	rās	(ġe)risen
	Proto-Germanic	⋆rīsō	⋆rais	⋆rizanaz
(S4) to steal	Modern English	steal	stole	stolen
	Old English	stele	stæl	stolen
	Proto-Germanic	⋆stelō	⋆stal	⋆stulanaz
(W1) to keep	Modern English	keep	kept	kept
	Old English	cēpe	cēpte	cēped
	Proto-Germanic	⋆kōpijō	⋆kōpidē	⋆kōpidaz
(W1) to send	Modern English	send	sent	sent
	Old English	sende	sende	(ġe)sended
	Proto-Germanic	⋆sandijō	⋆sandidē	⋆sandidaz
(W2) to love	Modern English	love	loved	loved
	Old English	lufie	lufode	(ġe)lufod
	Proto-Germanic	⋆lubō	⋆lubōdē	⋆lubōdaz
(W3) to like	Modern English	like	liked	liked
	Old English	līcie	līcode	(ġe)līcod
	Proto-Germanic	⋆līkijō	⋆līkdē	⋆līkdaz

Table 2.3 English verbal inflection classes explained diachronically. "*" stands for reconstructed forms.

Person, Number	**Future**	**Present**	**Past**	**Imperative**
1SG	*napišu*	–	M *napisal* F *napisala*	–
2SG	*napišeš`*	–	M *napisal* F *napisala*	*napiši*
3SG	*napišet*	–	M *napisal* F *napisala* N *napisalo*	–
1PL	*napišem*	–	*napisali*	–
2PL	*napišete*	–	*napisali*	*napišite*
3PL	*napišut*	–	*napisali*	–
PARTICIPLE, PAST, ACT.			*napisavšij*	
PARTICIPLE, PAST, PASS.			*napisannyj*	
PARTICIPLE, PAST, ADV.			*napisav(ši)*	

Table 2.4 An example of inflectional paradigm. Conjugation of Russian word `napisat`` "write (perfective)".

such as present tense indicative and participle, present tense. Contrary to one of the most common claims about inflectional paradigms, namely their completeness (non-defectiveness), we indeed observe that here it does not hold true. Certainly, one could argue that they should be treated as different paradigms. Now, if we compare future tense forms, we notice that their formation processes differ as well. The imperfective forms are produced by applying `be.Person.Number + infinitive` pattern. Finally, participle forms are considered to be a part of the paradigm while 1) the part of speech changes; and 2) each of them has its own distinct paradigm.[10] This means that participles are behaving similar to derivation and are, indeed, borderline cases as discussed later.

Adjective declension system might also present empty slots or variations in forms. As Table 2.6 shows, the Russian instrumental case allows two different endings for feminine forms. Also, the short form of the adjective allows variation in stress for feminine and neuter. In addition, Russian grammar assigns restrictions to the short forms, i.e. only can qualitative adjectives have short forms. For the rest of adjectives, such as `derevjannyj` "wooden" or `anglijskij` "english", the corresponding slots will be empty. Similar to

[10]Therefore, it is crucial for any NLP system to capture aspect well.

Person, Number	Future	Present	Past	Imperative
1SG	*budu pisat`*	*pišu*	M *pisal* F *pisala*	–
2SG	*budeš` pisat`*	*pišeš`*	M *pisal* F *pisala*	*piši*
3SG	*budet pisat`*	*pišet*	M *pisal* F *pisala* N *pisalo*	–
1PL	*budem pisat`*	*pišem*	*pisali*	–
2PL	*budete pisat`*	*pišete*	*pisali*	*pišite*
3PL	*budut pisat`*	*pišut*	*pisali*	–
PARTICIPLE, PRESENT, ACT.		*pišuščij*		
PARTICIPLE, PAST, ACT.		*pisavšij*		
PARTICIPLE, PAST, PASS.		*pisannyj*		
PARTICIPLE, PAST, ADV.		*pisav(ši)*		

Table 2.5 An example of inflectional paradigm. Conjugation of Russian word `pisat`` "write (imperfective)".

English, Russian adjectives have comparative forms that are expressed morphologically or non-morphologically. Typically they are not included into the same paradigm.

Szymanek (2010) provides an example for declension of Polish adjective `zły` "bad, evil" and noun `zło` "badness, evil". Although they share the same stem `zł`, there are two

Case		Masculine	Neuter	Feminine	Plural
NOM		*krasnyj*	*krasnoje*	*krasnaja*	*krasnyje*
GEN		*krasnogo*		*krasnoj*	*krasnyx*
DAT		*krasnomu*		*krasnoj*	*krasnym*
ACC	AN	*krasnogo*	*krasnoje*	*krasnuju*	*krasnyx*
	INAN	*krasnyj*			*krasnyje*
INS		*krasnym*		*krasnoj(-oju)*	*krasnymi*
PREP		*krasnom*		*krasnuju*	*krasnyx*
Short Form		*krasen*	*krasno*	*krasna*	*krasny*

Table 2.6 An example of inflectional paradigm. Declension of Russian word `krasnyj` "red".

distinct sets of inflectional suffixes used, and therefore, the two words belong to two different paradigms. Therefore, the change of paradigm typically yields the change of word-class.[11]

2.2.4.2 Theoretical approaches to Inflections

In this thesis, we discuss approaches to generation of word forms. Therefore, we provide a brief summary of linguistic views that address the following questions. First, which forms are likely to be memorised and serve as a basis for generating the others? Second, whether we need to model processes underlying the word production (agnostic to morphemes) or we actually need to identify and arrange morphemes.

Regarding the first question, as we mentioned earlier, people speaking languages such as Latin or Sanskrit are unlikely to memorise all word forms. Some might argue that both exhibit a modest number of possible forms, but cases of languages with higher degree of synthetism add more counter-evidence of the memorisation hypothesis. In Archi, a Caucasian language (Kibrik, 1998), the number of forms for every verb can reach 1.5 million. Another counter-argument is humans' ability to produce and comprehend forms they did not observe before. Numerous experiments on child language acquisition (e.g. Ervin, 1964) provide strong support for the idea that the forms are actually generated rather than memorised. This is related to computation and storage problem. Various psycholinguistic studies such as Jaeger et al. (1996) and Ullman (2001, 2004) claim that regular inflectional forms are processed as rules rather than stored in memory while irregulars and derived words are part of the mental lexicon and are retrieved as a whole. Although Baayen (2007) argues that such a dichotomy is overly simplistic. There is also general agreement in linguistics that frequently used complex words become part of the lexicon as wholes, while most other words are likely to be constructed from constituents (Aronoff and Lindsay, 2014; Bauer, 2001).[12] These words typically follow derivational patterns, or rules, such as adding `-able` to express potential or ability or applying `-ly` to convert adjectives into adverbs. This idea

[11] The last argument could be easily questioned by a counter-example such as Russian nouns `škola` "school" and `škol'nik` "school boy" where we do not observe a change of word-class.

[12] Although some neuroscience studies such as Kireev et al. (2015) reported that (Russian) regular verbs production does not involve Broca's area while for irregulars it is required in order to choose the right paradigm and rule of production.

is also in agreement with Chomsky's Universal Grammar principles, and has been realised in the form of **Dual Mechanism Theory**.[13] The theory allows mechanisms for extraction of grammatical rules by the systematic analysis of inputs. Pinker and Prince (1988) call them "symbol-manipulating processes". The proponents of the Dual Mechanism rely on the studies of aphasia and other language disorders. For instance, Marslen-Wilson and Tyler (1997) witnessed a dissociation between regular and irregular English past tense production in their study with aphasics whose native language is English. These differences has been confirmed in further studies with children having Williams syndrome (Clahsen and Almazan, 1998). The difference also applies to processing of semantically and syntactically anomalous constructions (Clahsen, 1999; Kutas and Hillyard, 1980). The two different approaches to the language faculty led to the famous *"The Past Tense Debate"* that is described in more detail in Section 2.3.5.1.

Hereafter, we base our models on the assumption that some word forms are actually generated on the basis of others. Next, we need to identify which forms are prior to others. Albright (2002) focuses on paradigms and investigates a hypothesis that speakers choose a single form as the base one, and it should be a surface form. As he notes, the base form should be "maximally informative", i.e., the one "that suffers the least serious phonological and morphological neutralizations" (Albright, 2002, p. 7). He further makes a stronger claim that this choice is global, i.e. all lexical items are produced on the basis of the same paradigmatic slot. Several models consider inflected form derivation from a *single* base form which typically corresponds to lemma form (for instance, for nouns it could be the one corresponding to the nominative singular).

In order to address the second question, recall two Hockett's models discussed in Section 2.2.1, "Item and Arrangement" and "Item and Process". "Item and Arrangement" requires knowledge of morphological rules and morpheme lists, although does not present any hierarchy of the forms themselves. The morphological rules operate on underlying representations of sub-parts of the words. The resulting representations are afterwards combined and transformed into surface forms. Distributed Morphology (Halle and Marantz, 1994),

[13] It is also refferred as "Dual Route".

DATR (Evans and Gazdar, 1996), and Lieber's syntactic approach to morphology (Lieber, 1992) belong to this class.

In "Item and Process" models (Hockett, 1954) new words are produced on the basis of others by means of morphological rules that are applied directly to their surface forms (sequentially). As Hockett notes, the past tense form `baked` is formed from `bake` by a suffixation process. Importantly, under this model, one form becomes superior to others, and this poses a question about the nature of the priority, whether it is historically motivated or some other criteria such as information-based guide it. Aronoff's word-based model (Aronoff, 1976), extended word and paradigm model (Anderson, 1992), lexical relatedness model (Bochner, 2011), and whole word morphology (Ford et al., 1997) represent this type of model.

In Section 2.3 we will discuss contemporary models representing both classes and show the superiority of process-level modelling. To summarise, in this section we discussed the notion of inflectional paradigm and briefly summarised existing linguistic approaches to morphology. The next section provides some background on the second type of morphology that we aim to model in this thesis, derivational. We discuss its relationship to inflection, and give a quick summary of linguistic views on it.

2.2.5 Derivations

The second major topic of the thesis relates to word formation, and, in particular, derivations. Here we provide a brief summary of linguistic approaches to derivation and also make an attempt to place inflections and derivations on a single scale of productivity and specificity. Similarly to the example for inflections, we can identify several slots forming a derivational paradigm. For instance, the verb `to write` has the agentive nominalisation (`writer`), the result or process nominalisation (`writing`) and the "ability" adjectivisation (`writable`). On the one hand, there are consistent patterns associated with each derivational slot, i.e. agentives are often expressed by `-er` or `-or` suffixes. On the other hand, most of the proposed paradigm tables appear to be very sparse. For instance, a significant number of verbs do not have a derivational form corresponding to patient nominalisation (e.g., `write` or

Case	Masculine	Neuter	Feminine	Plural
NOM	*udobočitaemyi*	*udobočitaemoje*	*udobočitaemaja*	*udobočitaemyje*
GEN	*udobočitaemogo*		*udobočitaemoj*	*udobočitaemyx*
DAT	*udobočitaemomu*		*udobočitaemoj*	*udobočitaemym*
ACC AN	*udobočitaemogo*	*udobočitaemoje*	*udobočitaemuju*	*udobočitaemyx*
ACC INAN	*udobočitaemyj*			*udobočitaemyje*
INS	*udobočitaemym*		*udobočitaemoj(-oju)*	*udobočitaemymi*
PREP	*udobočitaemom*		*udobočitaemuju*	*udobočitaemyx*
Short Form	*udobočitaem*	*udobočitaemo*	*udobočitaema*	*udobočitaemy*

Table 2.7 An example of inflectional paradigm. Declension of Russian word udobočitaemyi "readable". "Short form" stands for a short forms of the adjective (when part of its ending can be truncated).

think). Moreover, the agentive nominalisation could also be expressed with less productive suffixes such as -ee as in standee, or -ist in cyclist. The last suffix class, at the same time, presents a whole set of other meanings sharing the concept person (compare artist, violinist, Baptist, capitalist, Marxist). While most attempts at morphological modelling have targeted inflectional morphology, derivational still remains largely unstudied.

2.2.5.1 Derivation and inflection: a continuous scale or a dichotomy?

Paradigms and word class change. Historically, neither structuralism (Harris, 1946) nor generative linguistics (Chomsky, 2014) made any clear distinction between inflection and derivation. Bloomfield (1933) noted that there were no obvious criteria for separating the two. Aronoff (1976) was one of the first to try to distinguish their characteristics.

Pairs such as sing – sings and read – readable differ in many ways, and, most importantly, each word of the last pair stands for a different lexeme. Read, as a verb, follows a particular conjugation scheme as well as readable, as an adjective, follows a particular declension pattern (consider its translation into Russian, udobočitaemyi as illustrated in Table 2.7). In this way, it seems to be quite clear that these pairs correspond to two different phenomena. But what happens in adverbalisation cases such as interesting – interestingly (or, in Russian, interesnyj – interesno)? In the case of adverbalisation, the meaning is quite predictive and the surface form realisation is regular (in English, it is usually just attachment of -ly suffix, in Russian it is also regularly expressed

by -o suffix). Unlike interesting, the form interestingly does not have its own paradigm. Therefore, should we consider it to be a part of the paradigm for interesting? Similar question raises towards adjective degrees. Should comparative and superlative forms be parts of another paradigm? In addition, linguists often highlight that inflections do not change word class (its part of speech) whereas derivations do. This does not hold true for participle formation as in move – moving. Even though participles change part of speech, they are usually considered as a part of a verb's inflectional paradigm.

Still, the "paradigm" concept is mainly applied and used in inflectional morphology, and is not specified outside of this domain. Historically, many language grammars were described as paradigms where a table slot corresponded to a particular feature combination. But if we face an undocumented polysynthetic language, it is troublesome to identify the features, and there is no straightforward way to find out what should form the paradigm.

Productivity. The distinction between the two types of morphology is also a terminological problem since no strict definition had been proposed in order to differentiate them. For instance, a criterion proposed by Aronoff states, "Inflectional morphology tends to be more productive than derivational morphology" (Aronoff and Fudeman, 2011, p. 169). This formulation places derivation and inflection into a continuous scale rather than identifies a binary criterion. The term *productivity* itself presents several senses, and, therefore, adds more vagueness. One of the most common senses relates to the regularity and states that the resulting word form is predictable *both* in its meaning *and* surface form. A good example for that is the English plural noun ending /z/. First of all, the rule only applies to English nouns with singular number property. Second, it is usually realised as adding -s or -es ending to the lemma form.[14] If we now look at derivational pattern -able/-ible with the meaning of "which can be Verb-ed", it attaches to almost all transitive verbs, i.e. it is productive and applies to a broad range of base forms, similarly to inflection. As (Bauer,

[14]On the other hand, there are some exceptional cases such as mouse - mice, child - children, or lemma - lemmata. Should these irregularities be considered as exceptions in inflections or attributed to derivations? Many of such irregularities can be addressed if we consider etymological information and look at the data diachronically rather than synchronically (as we presented in Section 2.2.4.1). Clearly, morpheme productivity changes over time. Two first cases originate from Old English whereas the last one comes from Ancient Greek, and, therefore, follows its plural formation pattern.

1988, pp.79-80) notes, "derivation is more productive than is generally thought, ... inflection is less productive than is frequently believed".

In both derivations and inflections we also observe competitive forms and idiosyncrasy. For instance, Booij (2006) gives an example of the English plural noun `brethren` that does not just signify a plural form of `brother` but rather refers to more specific concept of male members of a religious community. Derivations usually present more such idiosyncratic cases. Mathews in *"Word and Paradigm"* (Matthews, 1965) model excludes suffix realisation from productivity and only considers the meaning. We observe that inflections are indeed more productive and form paradigms, while derivations do not typically behave like that.

Restrictedness. It is hard to discuss productivity without mentioning restrictedness. These two concepts typically go side-by-side. Unlike inflections, many derivational patterns are subject to semantic, pragmatic, morphological or phonological restrictions. Consider the English patient suffix `-ee`, which cannot be attached to a base ending in `/i(:)/`, e.g., it cannot be attached to the verb `free` to form `*freeee`. Restrictedness is closely related to productivity, i.e., highly productive rules are less restricted. A parsimonious model of derivational morphology would describe forms using productive rules when possible but may store forms with highly restricted patterns directly as full lexical items.

Syntactic relevance. Addressing all these terminological issues with definition, in order to distinguish inflectional morphology, Anderson (1982, 1992) proposes to focus on the syntactic relevance of inflections, "Inflectional morphology is what is relevant to syntax" (Anderson, 1982, pp.587). Note that the definition is too broad, i.e., although change of part-of-speech is relevant to syntax, it is out of scope in Anderson's model. Anderson makes agreement the main focus of the distinction. If a language marks Subject–Verb agreement, a verb specifies its subject's properties (for instance, when it agrees in number and gender with a corresponding noun/pronoun) rather than its own properties. For Anderson, this is the main characteristic of inflections.

Inflection	Derivation
ASPECT	*Mode of Action*
PASSIVE VOICE	*Derived intransitive verbs*
PARTICIPLE	*Verbal Adjective*
INFINITIVE	*Verbal Noun*
CONCRETE CASE	*Denominal Adverb*
GENDER OF ADJECTIVE	*Gender of Substantive*
COMPARATIVE	*Nominal Intensive Forms*

Table 2.8 An example of inflection – derivation mapping from Kuryłowicz (1964).

Parallels between inflections and derivations. Kuryłowicz (1964) draws parallels between inflectional and derivational morphology. He refers to a change from an inflectional to a derivational category as lexicalisation, and the opposite process as grammaticalisation. Table 2.8 demonstrates some examples of how inflectional and derivational categories can be mapped. In order to illustrate this, Kuryłowicz gives the following examples, `Apes are intelligent animals`. Here, the word `apes` carries two meanings, the meaning of plural (which is expressed in its form and syntactic relations) and the notion of a collective noun, because the sentence refers to the whole species rather than some of them. By numerous examples, he shows that these transitions are continuous rather than binary and are clearly seen if we look at languages diachronically. If we look at passivisation processes, for instance, the transformation of the utterance `John.NOM wrote the book.ACC` from active to passive voice, i.e. `The book.NOM was written by John.INS`, that differs from the original sentence syntactically but not semantically, although the latter one also allows us to omit the agent of the action, i.e. `The book was written`.[15] Now, if we consider the utterance `the red light`, we can apply an analogical transformation and get `the redness of the light`. And in order to enable such a transformational process one needs to allow abstract noun formation on the basis of adjectives. The relation between the former and the latter sentences in this case is similar to active-to-passive transformation.[16]

[15] Kuryłowicz additionally argues that the binary construction initially (in Old Latin, Arabic) was primary with respect to the tripartite expression and might have been the main function of the passive construction.

[16] Another example comes from deverbal abstracts such as `The flower blossoms`→`The blossoming of the flower`.

Relevance to stem and morpheme order. Finally, one of the Greenberg universals (Greenberg, 1963) postulates that derivational suffixes typically occur closer to the stem than inflectional suffixes. And it is commonly accepted that inflections are more relevant to syntax, and, therefore, appear closer to the word's boundaries. As Baker points out, "morphological derivations must directly reflect syntactic derivations (and vice versa)" (Baker, 1985, p. 375), a statement known as *the Mirror Principle*. Rice (2000) questions this point by providing some examples from Athapaskan languages, where inflections and derivations appear in less fixed order. A similar problem raises if we look at the ordering of derivational suffixes. Why are some particular combinations and orderings more accepted than others? As Saarinen and Hay (2014) note, why is `reddishness` better than `rednessish`? Hay and Baayen (2002) and Hay and Plag (2004) investigate possible restrictions on affix combinations in English derivation and propose the "complexity based ordering" model which suggests that suffixes that are less likely to be detached appear closer to the word stem than those that can be easily removed. In other words, the suffixes that are more relevant to the word stem appear closer to it (in accordance with Greenberg universals). In a more recent study, the authors also test the model on a larger set of suffixes and find that the suffix proximity to stem correlates inversely with its productivity (Plag and Baayen, 2009). By looking at Bantu languages, Hyman (2003) demonstrates that affix systems vary, and there are always two competing objectives: one forces the affix to be compositional and another one pushes it to follow a fixed order.

In this thesis, we will model and discuss paradigmaticity of both inflections and derivations and attempt to map them on a continuous scale from more prototypical and compositional instances to idiosyncratic ones (Booij, 2006; Bybee, 1985; Dressler, 1989; Scalise, 1988).

2.2.5.2 Problems in studies of derivations

There are general methodological issues researchers often face when attempting to study derivations. Since derivation presents a less regular and systematic structure than inflection, it leads to problems related to its productivity and transparency.

Many researchers rely on analyses of data done by themselves based on their own knowledge of a language. Such self-generated data has a range of flaws such as biases towards more frequently used and familiar items, which might have lexicalised meanings, and therefore lead to a conclusion that derivations present more idiosyncratic meanings than they really do. One way to address such a problem is to focus on neologisms since they are often more morphologically transparent. Neologisms, typically being low-frequency, are less likely to come to mind. Some researchers generate neologisms themselves, and afterwards assess the words as existing or non-existing based on their own intuition. But personal intuition might not be a reliable source. Lieber (1980) stated that "`Re-` and `un-` could only be attached to verbs invoving a change of state, and `kill` is not such a verb". Later, in Lieber and Štekauer (2014) she notes that corpus analysis (on the base of Corpus of Contemporary American English) shows a few occurrences of `rekill` and `unkill`. Although both of them are infrequent, they are still possible in appropriate contexts. As Aronoff notes, "Though many things are possible in morphology, some are more possible than others" (Aronoff, 1976, p.35). This leads to two main conclusions. First, we should not rely on our own intuition but rather use large corpora. Second, possible words should not be presented in isolation, they need an example of an appropriate context. Usage of corpora statistics additionally allows for a better representation of derivational pattern distributions.

2.2.5.3 Derivational Paradigm

As Plank (1994) notes, inflections form a comparatively closed system (paradigms), while derivations are not that well-organised in terms of form-sense regularities. Dressler (1989) argues that in both categories we observe more or less prototypical instances. This is supported by Stump (1991) who suggests that arguments motivating inflectional paradigms could be also applied to derivations and, therefore, allow the possibility of viewing derivations paradigmatically. Derivational processes may be organised into paradigms, with slots corresponding to more abstract lexico-semantic categories for an associated part of speech (Booij, 2008; Corbin, 1987; Štekauer, 2014). Lieber (2004) presents one of the first theoretical frameworks to enumerate a set of derivational paradigm slots, motivated by previous studies of semantic

primitives by Wierzbicka (1988). Contrary to that, Marie (1994) says that derivations should be regarded in a fundamentally different way than inflections.

A key difficulty comes from the fact that the mapping between semantics and suffixes is not always clean; Lieber (2004) points out the category AGENT could be expressed by the suffix `-er` (as in `runner`) or by `-ee` (as in `escapee`). However, both `-er` and `-ee` may have the PATIENT role; consider `burner` "a cheap phone intended to be disposed of, i.e. burned" and `employee` "one being employed", respectively. Potential nominalisations corresponding to the RESULT of a verb could be `-ion`, `-al` and `-ment` (Jackendoff, 1975). Although typically an English verb employs only a single suffix realisation (`refuse`↦`refusal` blocks other candidates such as *`refusion` and *`refusement`), there are cases such as `deportation` and `deportment` but here they signify two *different* concepts. It is still unclear what the derivational paradigm should contain and which categories should fill the slots. Comparatively little work has been done in this area.

Lexeme-Morpheme Based Morphology (LMBM) One promising theoretical framework describing derivational relations, *Lexeme-Morpheme Based Morphology (LMBM)*, was proposed by Beard (1995). There, he separates lexical morphemes (noun, adjective, and verb stems) from grammatical ones. The LMBM theory was largely inspired by earlier ideas of parallels in lexical (semantic) and syntactic (inflectional) derivational processes expressed in works by Jerzy Kuryłowicz and Aleksandar Belić.

Kuryłowicz (1936) discussed the functional distinction between lexical (semantic) and syntactic derivations. The study was inspired by recent investigation on duality in parts of speech by Slotty (1932). Slotty noted that nouns refer to objects and simultaneously play the syntactic role of Subject or Object. Similarly, an adjective signifies some object's quality and plays a modifier role. A verb refers to an action or a state or a process and functions as a predicate. On the other hand, this does not always hold true, and we can come up with multiple counter-examples. Consider, the following sentences in Russian: `Belaja sobaka bejit` "A white dog is running" and `Sobaka bela` "A dog is white". Both sentences contain an object's quality ("white"), but in the first one it is clearly a modifier

whereas in the second one it plays the role of predicate. Note that in the second sentence the Russian form for "white" changed from the full form to the short one.[17] Based on these observations, Kuryłowicz identifies primary and secondary syntactic functions and suggests a rule of formal transformation that states the following principle, "If changes in syntactic function of word A lead to the change of its form from A to B; then the initial form should be considered as a primary and the derived one as a secondary". As an example, he compares the Latin *amat* "love.V.PRES.3SG" and amans "Loving.PART.M.SG.NOM" that are only different syntactically. Since amans is derived from amat, we conclude that for words of action the predicate function is the primary one, and modifier is the secondary. So, he formulates the notion of syntactic derivation as a form with the same lexical (semantic) content as in the initial form, but different syntactic function. Contrary to that, lexical (semantic) derivations do not change syntactic function. Functions of diminutive nouns are the same as their base nouns, and transformation of a verb from imperfective to perfective form does not affect its syntactic role. By extending the definition of derivation, we can also re-define inflection. For instance, plurality is relevant to both syntactic (as it participates in agreement) and lexical (as it refers to the number of objects) roles. The Russian example of "white dog" also demonstrates that in Indo-European languages adjectives in the predicate role are often marked by an auxiliary verb ("be" + adjective).

Beard's LMBM theory stands on three hypotheses:

- *The Separation Hypothesis* distinguishing inflectional and derivational processes from their phonological realisation (making the functions independent of their realisation). The hypothesis addresses one-to-many and many-to-one mapping between a morpheme and its function as well as allows realisation without any functional variation (empty morphemes such as −al in dramatical, syntactical);

[17] Unlike the full form, the short does not inflect for the case. Short forms can only be formed from qualitative adjectives, but (somewhat surprisingly) historically are superior to the full ones since initially the Proto-Slavic language only had short adjective forms that were of the same nature as verbs.

- *The Unitary Grammatical Function Hypothesis* proposing a one-to-one mapping between inflectional and derivational categories (e.g. Agent, Patient, Location, Possession; 44 categories in total);

- *The Base Rule Hypothesis* that states that the universal categories of a word must originate in a base component (contrary to a transformational component in the sense of Chomsky (2014)).

Beard identified four groups depending on functions, inherent feature expression, expressivity, and part of speech change. In terms of the functional derivation system, he proposed the idea of mapping derivational slots into a grammatical case: "It, therefore, seems more likely that this type of derivation is based on case functions: (nominative of) subject, (accusative of) object, (locative of) place (`bakery`, `fishery`), (genitive of) possession (`dirty`, `forested`) and material (`oaken`, `woolen`), (ablative of) origin (`American`, `Brazilian`), (dative of) purpose (`roofing`, `siding`), (instrumental of) means (`cutter`, `defoliant`)" (Beard, 2017, p. 59).

Featural derivation does not change the part of speech of the base form and only affects some inherent feature of the word such as gender. Jakobson (1932) (and more in Jakobson (2011)) studied markedness of gender in Russian. He noted that most nouns attribute masculine as the default one, and, therefore, it is not marked. Feminine forms are typically marked by an extra suffix such as `-k` (`student`, "student.M" → `studentka`, "student.F"), `-sh` (`kassir`, "cashier.M"→`kassirša`, "cashier.F"), `-ess` (`poet`, "poet.M"→`poetessa`, "poet.F"), and others. These examples can be opposed to the pure masculine cases like `brat`, "brother.M" and `otec`, "father.M".

Transposition is another axis of viewing derivations. It reflects a simple change of category without any functional change. For instance, `walk` → `walking` (V → N) or `new` → `newness` (A→N).

Finally, *expressive* derivation expresses the speaker's opinion about the object (e.g., diminutives). It neither changes the referential scope, nor the lexical category of the base. As an example consider the three grades of the Russian word for "rain": `dožd`` ("rain"),

dožd*ik* (“little rain”), and dožd*iček* (“very little rain”), all referring to the same conceptual category. One might also add dožd*išče* (“strong rain”).

To conclude, we briefly summarised existing views in linguistics on inflectional and derivational processes and outlined various parallels between them. The following two sections (Sections 2.2.6 and 2.2.7) are devoted to the datasets that we use and tasks we target in this thesis.

2.2.6 Resources

Here we list the resources that we use in experiments that are discussed in Chapters 3, 4 and 5.

NOMLEX NomLEX (Meyers et al., 2004) is a dictionary of nominalisations in English. The database relates the nominal complements to the argument structure of the corresponding base verb. In total there are about 1,000 entries. Note that the dataset mainly targets syntactically motivated derivations and does not contain information on subtle semantic differences such as diminutive or feminitive forms.

```
(nom      :orth "promotion"
          :verb "promote"
          :nom-type((verb-nom))
          :verb-subj ((n-n-mod) (det-poss))
          :verb-subc ((nom-np :object ((det-poss)(n-n-mod)(pp-of)))
                      (nom-np-as-np :object ((det-poss) (pp-of)))
                      (nom-possing :nom-subc ((p-possing :pval ("of"))))
                      (nom-np-pp :object ((det-poss) (n-n-mod) (pp-of))
                                 :pval ("into" "from" "for" "to"))
                      (nom-np-pp-pp :object ((det-poss) (n-n-mod) (pp-of))
                                    :pval ("for" "into" "to") :pval2 ("from"))))
```

Figure 2.1 An example of the NOMLEX data entry for promotion.

CELEX CELEX (Baayen et al., 1993) is a database that contains inflectional and derivational morphology information for English, German, and Dutch. Derivational forms are provided as segmentation with POS specification, e.g. attract.VERB + -ion.NOUN→ attraction.NOUN. For English there are approximately 53,000 entries. In addition, it provides approximately 161,000 inflected forms with patterns describing the word change process.

UniMorph: universal morphological annotation schema Most investigations on morphology modelling focused on high-resource languages such as English, German, Russian, or French. Such a view is biased towards languages with modest levels of morphological complexity and a large amount of data available, including grammars and lexicons. Luckily, during the last decade, a lot of information on languages has appeared in open source resources such as Wiktionary. Although the information is represented in many languages, the format of annotation is often quite inconsistent between languages, meaning it has limitations as a parallel resource. These factors motivated the creation of a universal language-independent schema for morphological annotation as well as a single repository called UniMorph (Sylak-Glassman et al., 2015a,b) and the organisation of a series of shared tasks to promote the development of systems for morphological analysis and inflection. The UniMorph schema aims to encode inflectional morphological meanings across the world's languages. The morphological features are represented as key-value pairs. For instance, "`pojedu [lemma=pojehat`, POS=VERB, Mood=INDICATIVE, Tense=FUTURE, Person=1, Number=SINGULAR]`" (I will go).[18] Here, each inflected form is provided with its lemma and a set of morphological features. The database contains inflectional paradigms for over 100 languages as of 20/08/2018.

2.2.7 Tasks

Before starting to describe the models, we first list a number of established tasks in morphology.

2.2.7.1 Inflectional Morphology

Lemmatisation is canonicalisation of the wordform, i.e. transformation of an inflected form into its dictionary variant (=lemma). Typically it also involves prediction of the morphological tags of the inflected form. For instance, the output of the lemmatisation of the form `created` will be its lemma form `create` and a set of tags such as `POS=VERB, Tense=Past`.

[18] The process of paradigm tables extraction from Wiktionary is described in Kirov et al. (2016).

Morphological Inflection goes the other way around, i.e. given a lemma form and the target tags the system has to predict the inflected one. Taking the previous example, `lemma=create, POS=VERB, Tense=Past` → `created`. The task can be easily generalised and formulated as **reinflection**, i.e. generation of one inflected form from another one of the same lemma. Typically, the system is provided with a source inflected form together with its tags and target form tags. For instance,
`form=creates, srcPOS=VERB, srcTense=PRES, srcNumber=SINGULAR, srcPerson=3, tgtPOS=VEBR, tgtTense=PAST` → `created`.
In more challenging scenario source tags might be omitted:
`form=creates, srcPOS=VERB, srcTense=PRES, srcNumber=SINGULAR, srcPerson=3` → `created`.
To some extent, it mimics an L1 language acquisition scenario. The reinflection task involves analysis of an inflected form followed by synthesis of a different inflection of the same form. This task supports a more realistic setting, wherein systems might not observe full paradigms.

Paradigm Completion Unlike the previous task on morphological reinflection, here the systems must complete the inflectional paradigm, i.e. they are supplied with a lemma and some of its inflected forms and asked to predict the remaining forms in the paradigm. Table 2.9 illustrates the data setting. Importantly, during training the systems are provided with *complete* paradigms. The task is also important in terms of measuring systems' ability to extrapolate. It reconstructs the condition of human learners when only a few complete paradigms are observed, and the learner has to generalise them to unseen ones. Certainly, such a task also assumes availability of dictionaries and mimics to some extent a typical L2 (second language) learning setting. The task also simulates learning from a very limited amount of data (restricted number of inflection tables).

2.2.7.2 Derivational Morphology

Due to the complications mentioned in Section 2.2.5.2, much less work has been done in terms of derivational morphology.

Lemma	Inflections	Inflection Tags	Lemma	Inflections
lěto	*lěta*	NOUN;NOM;PL	*śěło*	–
lěto	*lětow*	NOUN;GEN;PL	*śěło*	–
lěto	*lěta*	NOUN;GEN;SG	*śěło*	–
lěto	*lěto*	NOUN;ACC;SG	*śěło*	–
lěto	*lěśe*	NOUN;ACC;DU	*śěło*	–
lěto	*lětami*	NOUN;INS;PL	*śěło*	–
lěto	*lětach*	NOUN;ESS;PL	*śěło*	–
lěto	*lěśe*	NOUN;ESS;SG	*śěło*	–
lěto	*lětoma*	NOUN;DAT;DU	*śěło*	*śěłoma*
lěto	*lěto*	NOUN;NOM;SG	*śěło*	*śěło*
lěto	*lětoma*	NOUN;ESS;DU	*śěło*	–
lěto	*lětom*	NOUN;INS;SG	*śěło*	–
lěto	*lětoju*	NOUN;DAT;SG	*śěło*	–
lěto	*lěta*	NOUN;ACC;PL	*śěło*	–
lěto	*lěśe*	NOUN;NOM;DU	*śěło*	–
lěto	*lětowu*	NOUN;GEN;DU	*śěło*	–
lěto	*lětoma*	NOUN;INS;DU	*śěło*	–
lěto	*lětam*	NOUN;DAT;PL	*śěło*	*śěłam*

Table 2.9 An example of training and test data for paradigm completion subtask of SIGMORPHON 2017. For training, systems are provided with complete paradigms such as the one for lěto, and, at test phase, systems are required to predict the missing forms (marked here by "–") as illustrated for śěło.

Paradigm Completion Similarly to inflections, we can formulate a task of paradigm completion for derivation. Here, provided with a possible derivational paradigm, we aim to predict a form for a target slot such as `base=run, POS=NOUN, Sense=Agent` → `runner`.

In addition, in the thesis we introduce contextual inflection and derivation which is a more realistic setting, i.e. we replace a morphological tag with sentential context and train a model to predict the word form that fits this context.

2.3 Modelling

Generally, we can identify three classes of morphological models, namely linguistically-inspired, FSA-based, and neural ones. This section provides a history of computational morphological models, starting from finite-state automata and finishing with state-of-the-art neural approaches.

2.3.1 Finite-State Machines

Mathematical Background Earlier approaches to morphology modelling (to address lemmatisation, inflection, and related tasks) used **finite-state machines**. In this section, we provide some theoretical background on this topic. A **deterministic finite-state machine (acceptor, Figure 2.2a)** is defined as a quintuple $A = (\Sigma, Q, q_0, E, F)$, where Σ is the input alphabet, S is a finite set of states, $q_0 \in Q$ denotes the initial state, $E : Q \times (\Sigma \cup \varepsilon) \rightarrow Q$ is the state-transition function; and $F \subseteq Q$ is the (possibly empty) set of final states. A transition $t = (p[t], l[t], n[t]) \in E$ between $p[t]$ as the previous state and $n[t]$ as the next state is labelled with $l[t]$.

The unweighted FSA provides us with boolean values, and, therefore, we can say that it is defined over the boolean semiring denoted as $(\{0,1\}, \vee, \wedge, 0, 1)$. The resulting value discriminates forms that are valid (accepted by FSA) from those that are not. In some cases it might be useful to output a score (*how valid* is the form?). Therefore, following the

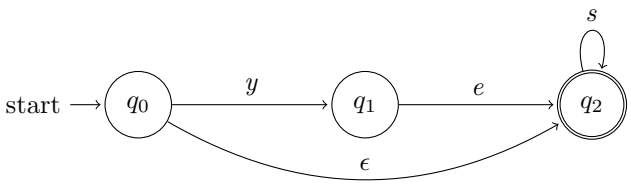
notation used in Mohri et al. (2002)[19], we define a **weighted finite-state acceptor (WFSTA, Figure 2.2b)** by adding extra parameters to the quintuple, namely λ and ρ, as well as enriching the transition t with transition weights $w[t]$: $t = (p[t], l[t], w[t], n[t])$. Importantly, we define WFSTA over the semiring $(\mathbb{K}, \oplus, \otimes, \overline{0}, \overline{1})$[20], usually the probability semiring $(\mathbb{R}, +, \cdot, 0, 1)$. A sequence of consecutive transitions $t_1 ... t_k$ such that $n[t_i] = p[t_{i+1}]$ is called a path in A. A successful path $\pi = t_1 ... t_k$ is a path starting at the initial state q_0 and ending at a final state $f \in F$. A label of the path $l[\pi]$ is a concatenation of individual arc's labels on this path. Its corresponding weight is evaluated as $w[\pi] = \lambda \otimes w[t_1] \otimes ... \otimes w[t_k] \otimes \rho(n[t_k])$. Finally, we say that a symbolic sequence x is recognised, or accepted by A, iff there exists a successful path π such that $l[\pi] = x$. The final weight assigned to x by A is then $\oplus$-sum over all successful paths π labelled with x.

A **finite-state transducer (FST, Figure 2.2c)** is an automaton whose transitions between states are labelled with both input and output symbols. That is, a path maps an input sequence to an output sequence. More specifically, we define the **weighted FST (Figure 2.2d)** as $T = (\Sigma, \Omega, Q, q_0, E, F)$ over the semi-ring $\mathbb{K}$, where Ω is an output alphabet, $E : Q \times (\Sigma \cup \varepsilon) \times (\Omega \cup \varepsilon) \to Q$. The transition t is enriched with an output label $l_o[t]$: $t = (p[t], l_i[t], l_o[t], w[t], n[t]) \in E$.

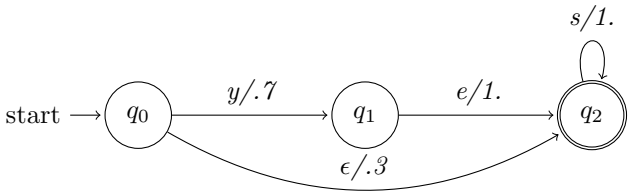
FST-based Morphology Figure 2.3 provides an example of FSA that recognises English adjective forms such as `cool, cooler, coolest, coolly` or `real, unreal, really`. The attentive reader might notice that it also accepts nonce words such as `unbigly` and `greenly`. Therefore, Antworth (1991) proposed a modification as illustrated on Figure 2.4. There, indices in *adj_root* signify different classes of adjectival stems. Note that this is a form-based approach, i.e. the transitions between the states rely on surface realisation rather than a semantic class. For instance, the negation prefix `un-` does not attach to adjectives denoting colors such as `red` or `blue`. The closest corresponding meanings

[19] See also Chris Dyer's notes on semirings from `http://demo.clab.cs.cmu.edu/fa2015-11711/images/6/63/Semirings.pdf`

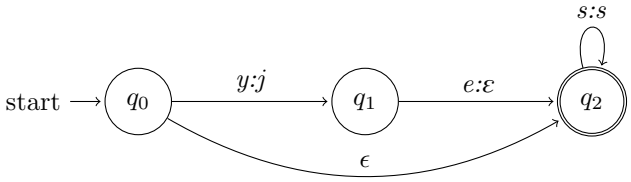
[20] Here, $\mathbb{K}$ is a set; $\oplus$ is an additive operator that is commutative and associative; $\otimes$ is a multiplication operator that is associative; $\overline{0} \in \mathbb{K}$ is an additive identity element, i.e. $\overline{0} \oplus a = a$, and an annihilator, $\overline{0} \otimes a = \overline{0}$; finally, $\overline{1} \in \mathbb{K}$ is a multiplicative identity element, i.e. $\overline{0} \otimes a = a$



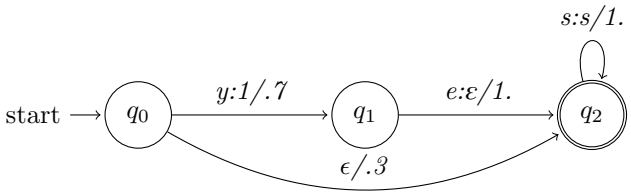
(a) Unweighted Acceptor.



(b) Weighted Acceptor.



(c) Unweighted Transducer.



(d) Weighted Transducer.

Figure 2.2 Finite-State Machines

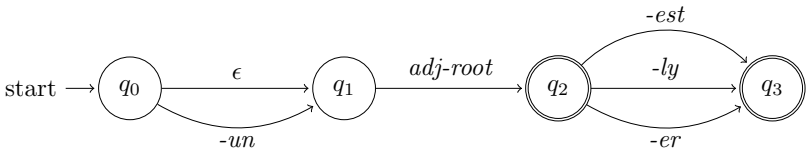


Figure 2.3 An FSA for English adjective morphology.

which denote absense of the color would more likely realised by means of `-less` suffix, i.e. `redless` or `blueless`.

As illustrated in Figure 2.5, FSAs can also be used to model derivational transformations in English. For instance, it shows the well-known fact that a verb ending with `-ise` can take

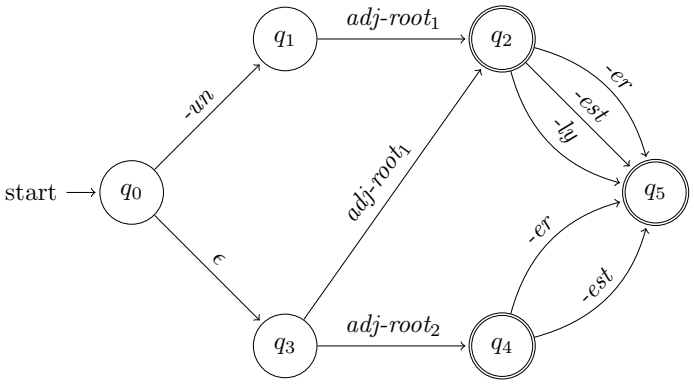


Figure 2.4 An improved FSA for English adjective morphology proposed in Antworth (1991).

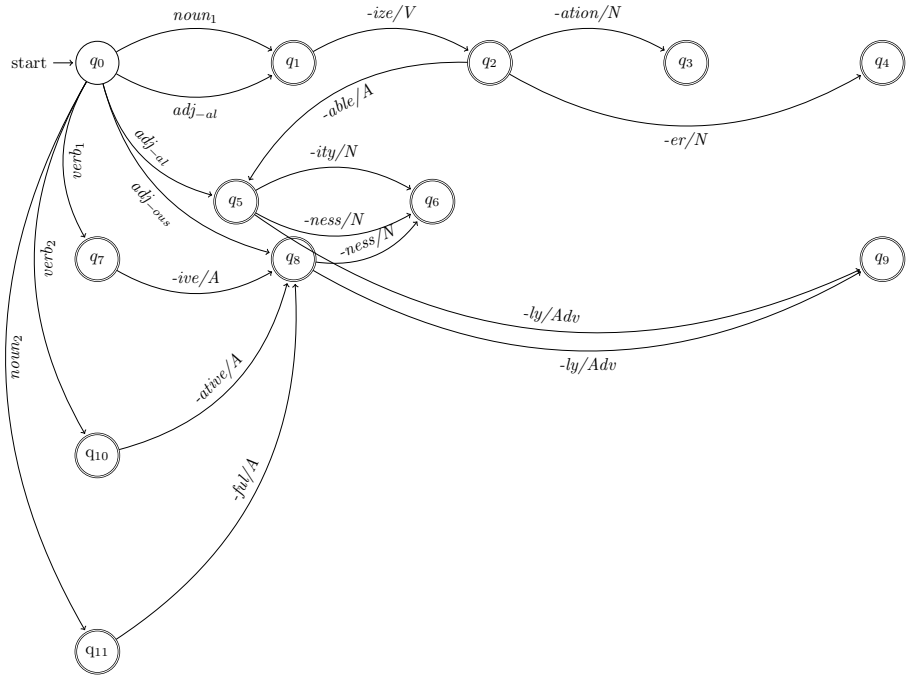


Figure 2.5 An FSA for English derivational morphology proposed by (Bauer, 1983; Porter, 1980; Sproat and Fujimura, 1993).

the nominalising suffix `-ation` (Bauer, 1983; Sproat and Fujimura, 1993). Using this FSA we can make predictions about the possible derived forms, e.g. `memory`→`memorise` → `((memorisation; memoriser); (memorisable` → `memorisability))`, although exceptions still apply, e.g. `special` → `specialise` → `((specialisation; *specialiser); *specialisable` → `*specialisability))`.

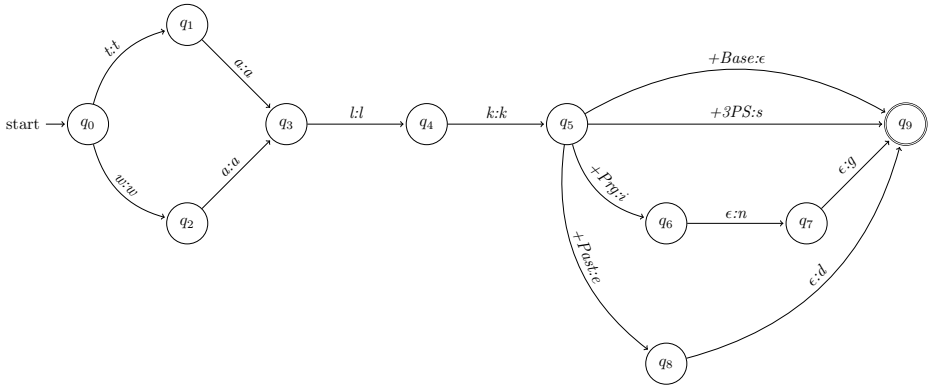


Figure 2.6 An FST describing regular English verbal inflection cases.

Two-Level Morphology FSTs and FSAs are the most popular and established methods for modelling inflectional morphology. The main assumption is that morphology can be represented with a set of connected states. Figure 2.6 provides an example for inflection. But this situation has not always been that optimistic. Until 1980s there was no established approach in computational linguistics to deal with complex morphological rules. In 1981 Koskenniemi, Karttunen, Kaplan and Kay started working on the morphological analysis problem. Kaplan and Kay (1994) showed that traditional phonological grammars formalised in Chomsky and Halle (1968) as ordered sequences of A→ B / [precontext _ postcontext] (a set of **rewrite rules** transforming abstract phonological representations into surface forms) describe regular expressions, and, therefore, by definition, could be represented by FSTs.[21] Earlier, Schützenberger (1961) proved the important property that for any pair of transducers applied sequentially there exists an equivalent single transducer. Given the aforementioned property, a cascade of rules could be represented by a single transducer without any intermediate representations. The idea is illustrated in Figures 2.7 and 2.8.

This important observation started an epoch of **two-level morphology**, the very first general model in computational linguistics for the analysis and generation of morphologically rich languages. The two levels, lexical and surface, are illustrated in Table 2.10.

There were many practical issues during two-level morphology development. First, the implementation of a compiler for rewrite rules turned out to be a hard task, since it required

[21] Such a notation describing pronunciation changes in different phonological and morphological contexts is usually referred to as an ordered set of context-sensitive rewriting rules.

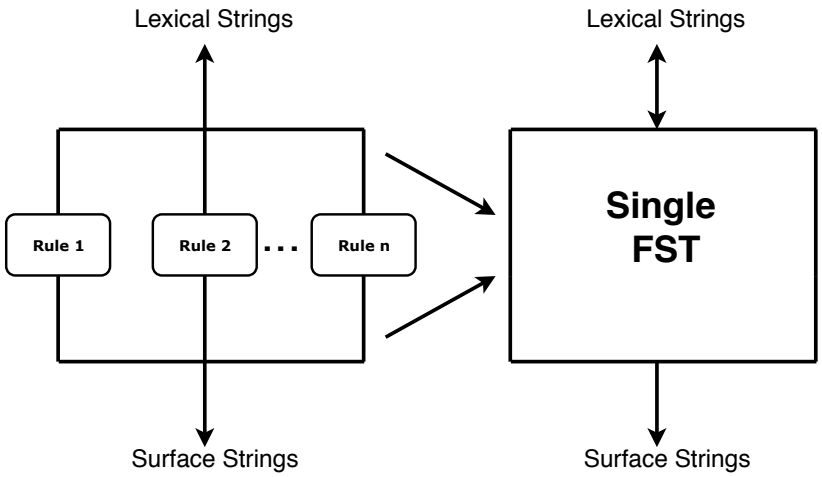


Figure 2.7 A cascade of rules mapped into a single FST (taken from Karttunen and Beesley (2001)).

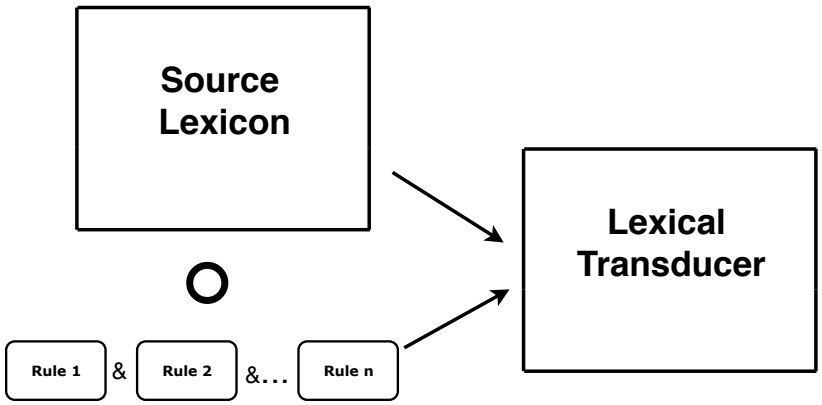


Figure 2.8 A lexicon intersected and composed with two-level rules (taken from Karttunen and Beesley (2001)).

m	o	v	e	+	e	d	Lexical
m	o	v	0	0	e	d	Surface

Table 2.10 An example of two-level morphology.

implementation of basic finite-state operations such as intersection, union, composition and complementation, which was quite challenging due to limited computational resources available at that time. Another problem was with morphological analysis. Rewrite rules describe only a one-directional process, generation, i.e. a mapping from lexical to surface forms. In this process, given that all operations are deterministic and obligatory, we are guaranteed to expect a single surface form. On the other hand, if we consider the morphological analysis

problem, the situation changes. There a single surface form might lead to various analyses, the number of which increases exponentially with the number of rules. The problem of analysis in Chomsky-Halle paradigms was challenging, and more difficult than the generation one. In Karttunen and Beesley (1992) the authors proposed to address it by FST formalisation of the lexicon as well, and composing the lexicon with the rules.

The two-level rules describe *regular relations*, similar to rewrite rules, and represent equal-length relations. This supports mapping lexical to surface strings and simulates an intersection of the automata. Unlike rewrite rules, the two-level rules are applied in *parallel*. The two-level architecture together with constraints on the lexicon overcomes the over-analysis issue. Importantly, the rules are symbol-to-symbol constraints rather than string-to-string relations as in rewrite rules.

To conclude, two-level morphology is based on the following ideas. First, rules are symbol-to-symbol constraints that are applied in parallel rather than sequentially as in rewrite rules. Second, the constraints can apply to the surface, to the lexical context, or both at the same time. Third, lexical lookup is performed in tandem with morphological analysis. The approach significantly advanced morphological analysis and generation for languages with concatenative morphology. A few years later, Beesley and Karttunen (2003); Cohen-Sygal and Wintner (2006); Kiraz (2000) proposed several modifications to address non-concatenation cases, such as root-and-pattern (templatic) morphology in Semitic languages.

2.3.2 Inflection as String-to-String Transduction and Automatic Learning of Edit Distances

As described in Section 2.2.7.1, in inflectional morphology we can identify two types of tasks, namely, **morphological inflection** and **lemmatisation**. The former takes an inflected form or a lemma together with morphological features of the target form and produces an inflection. The latter does the opposite, i.e. it produces a lemma (sometimes supplied with morphological feature values) from a given inflected form. Both belong to the task of string-to-string transduction. More specifically, we aim to learn a systematic mapping

between an input string x to an output string y. Such a task also comprises orthographic mapping between pronunciation and spelling, as well as cognates and loanword translation.

In many cases, a string transduction task requires evaluation of edit distance between strings x and y, i.e. the minimum number of insertions, deletions, substitutions needed to transform one string into another (Levenshtein, 1966). The distance between two strings $x^t \in A^T$ and $y^v \in B^V$ is defined as $d_c(x^t, y^v) = min \left\{ \begin{array}{l} c(x_t, y_v) + d_c(x^{t-1}, y^{v-1}), \\ c(x_t, \varepsilon) + d_c(x^{t-1}, y^v), \\ c(\varepsilon, y_v) + d_c(x^t, y^{v-1}) \end{array} \right\}$ where $d_c(\varepsilon, \varepsilon) = 0$, and $c(\cdot, \cdot)$ corresponds to cost of edit operation. The distance is computed in $O(TV)$ time using dynamic programming (Masek and Paterson, 1980). Bahl and Jelinek (1975) proposed a *stochastic* interpretation of edit distance, and Ristad and Yianilos (1998) introduced an algorithm to *automatically learn the edit costs* from a training corpus. The authors proposed two models, **Viterbi** and **stochastic edit distances**. The Viterbi method evaluates the most likely edit distance whereas the second one considers all possible ways to generate a string pair evaluating a joint probability of x^T, y^V.[22] This work was the first to show efficiency of the usage of stochastic models in pattern recognition. Cotterell et al. (2014) generalised Ristad and Yianilos (1998)'s stochastic edit distance model by enriching it with contextual features and expressing contextual probabilities by a conditional log-linear model. They also proposed to model the conditional probability $p(y|x)$ as a probabilistic finite-state transducer (PFST) because, unlike weighted FST, it does not require a separate calculation of normalizing constant Z_x for every x.[23] In addition, PFSTs are faster to compute gradients. They tested the models on spelling correction tasks (i.e. not related to morphology) without reliance on dictionaries or language models and presented an improvement over vanilla stochastic edit distance model (Ristad and Yianilos, 1998) in terms of log-likelihood.

[22]The costs of edit distances were estimated as the parameter of a memoryless stochastic FST. In order to solve this problem, the authors applied an expectation maximisation (EM) technique (Daum (1970); Dempster et al. (1977))

[23]Generally, WFSTs perform better than PFSTs in linguistic tasks (Dreyer et al., 2008).

2.3.3 Inflectional Paradigm Modelling

2.3.3.1 From FSTs to Neural Methods

Wicentowski and Yarowsky (2000) proposed a corpus-based algorithm for inducing rules of inflectional transformations between a lemma and its forms that accounts for frequency, contextual similarity, transduction probabilities and weighted string (Levenshtein) similarity. Wicentowki's model essentially learns and then applies replacement rules that are identified by deterministic alignment and segmentation. They focused only on English past prediction task and reported 99.2% accuracy on it (on both regular and irregular forms).

Dreyer et al. (2008) addressed (more general) morphological inflection and lemmatisation tasks and presented a conditional log-linear model which adds to initial character-level features two extra latent alignment ones. The authors propose to consider joint n-grams instead of independent modelling of each (x,y) aligned pair, the approach that has been shown to be useful before by Deligne and Bimbot (1995) and Bisani and Ney (2002).[24]

The task is stated as follows. Given two strings, an input x and an output y, we need to estimate a probability of $p(y|x)$. In order to do so, we estimate possible alignments between x and y and introduce two latent variables. The first latent variable l_1 encodes the word class, or a particular paradigm. For instance, `speak`, `break`, `steal` could be assigned to the same class since they follow the same form changing pattern. The second variable l_2 identifies numbered regions in the string pair in such a way that identity characters `(s,s)` belong to even regions (match) whereas regions with character change such as `(e,o)` fall in odd ones (deletion, insertion, substitution). This guides the model towards useful splits. Both latent variables are learned using the gradient-based optimisation technique L-BFGS from Liu and Nocedal (1989). Experiments using CELEX (Baayen et al., 1993) show that in most cases the model outperforms a character level SMT model trained with Moses (Koehn et al., 2007), and the major improvements come from latent variables. Most of the errors come from either

[24]Deligne and Bimbot (1995) attempted statistical language modelling task by jointly modelling a sentence and its possible segmentations expressed by n-grams and variable-length multigrams (over words), and showed superior perplexity results of the latter. Bisani and Ney (2002) used joint multigram models over sequences of phonemes and graphemes in order to perform grapheme-to-phoneme conversion and achieved reduction in error rates for English and German.

wrongly copying input characters to the output or applying regular paradigms to irregular forms (although the split on regular and irregular conjugation had been learned). For the lemmatisation task, the model was compared to a set of Wicentowski and Yarowsky (2000)'s models. The n-gram-based model augmented with latent variables performs substantially better than the baseline and leads to error reduction over several languages.

Rastogi et al. (2016) proposed a hybrid model that enriches FST with *neural* contextual representation. The paper addressed the problem of traditional FSTs which require manual design of the states and the features. In particular, for two states, h and h' connected with an arc that corresponds to the edit **s:t**, we need to assign an arc weight. The weight depends on the edit itself as well as the states. Essentially, h corresponds to the alignment of prefixes of two strings (parts preceding the current edit) and h' summarises the alignment of their suffixes (parts following the edit). Therefore, the **s:t** edit weight depends on context only through h and h'. The authors proposed to incorporate sequence-to-sequence neural model in order to represent context. More specifically, they incorporated the log-bilinear model from Salakhutdinov et al. (2007) to compute the edit function. The authors evaluated their model on morphological re-inflection and lemmatisation tasks and compared them against the Moses phrase-based MT (Koehn et al., 2007) and Dreyer et al. (2008)'s baselines. Using the CELEX dataset (Baayen et al., 1993) they demonstrated that their approach surpasses the Moses system and performs on par with the best setting from Dreyer et al. (2008).

Nicolai et al. (2015) studied morphological inflection task in paradigm-aware and agnostic settings. The authors described an inflectional model as discriminative string transduction where character-level operations are applied in order to transform a lemma with tags into forms. The main motivation for such a model comes from statistical machine translation into morphologically rich languages. In order to deal with data sparsity, Fraser et al. (2012) and Clifton and Sarkar (2011) proposed to convert forms in the target language into lemmata, and then during decoding make a prediction about their morphological tags and transform them into a proper inflected form. Unlike Durrett and DeNero (2013) and Hulden et al. (2014) that are discussed later in the section, Nicolai et al. (2015) do not aim to learn paradigm-wide regularities by doing multiple string alignment but rather align each form to its lemma.

Contrary to previous approaches, they do not concentrate on unchanged characters within a paradigm, but rather extract regular small multi-character operations observed across multiple paradigms.[25] Then from the aligned sequences they extract rules in such a way that each minimal multi-character transformation is assigned a separate rule. Once the rules are ready, a discriminative semi-Markov model adapted from Zens and Ney (2004) is applied in order to select a rule appropriate for a given lemma form. Experiments on the Wiktionary dataset taken from Durrett and DeNero (2013) show that their model outperforms that of Durrett and DeNero (2013) (discussed below) in the paradigm-agnostic setting, and performs on par when the systems are provided with complete paradigms.[26] Finally, error analysis shows that many mispredictions involve circumfixation, irregularities (over-correction of the forms) and are also related to difficulties in syllable and compound boundary identification in the case of Finnish.

As we note in Section 2.2.4.2, inflected forms should be considered together as a complete paradigm since they reinforce each other.[27] Earlier approaches mainly focused on modelling the probabilistic relationship between two strings. Dreyer and Eisner (2009) proposed building up joint models of three or more strings (partial paradigms). In their work, the authors introduced an undirected graphical model, in particular a Markov Random Field, in which the factors are the weighted FSTs introduced earlier.

Durrett and DeNero (2013) further develop the idea of joint form modelling and attempt to generate a complete paradigm table for a given lemma. First, they align the forms within a single paradigm by means of edit distance. In particular, they used a version of a dynamic (position-dependent) edit distance earlier proposed by Eisner (2002) and Oncina and Sebban (2006). There, insertion, deletion, and substitution are assigned a weight of 0, and match weight equals $-c_i$, where i is an index in the lemma form, and c_i is the number of forms for which there exists a match at the position i. This scheme encourages a lemma form to be aligned to all of its inflected forms. Next, they extract rules from aligned pairs by identifying

[25] In order to find the aligned pairs they use EM-based M2M aligner (Jiampojamarn et al., 2007)

[26] They additionally test the system's performance in a low-resource setting when it is provided with 50 or 100 paradigm tables The accuracy of the models is comparable with established baselines from Durrett and DeNero (2013) and Dreyer and Eisner (2011).

[27] The same motivation comes from cognate modelling and language re-construction.

changed spans, i.e. each rule corresponds to a sequence of edit operations needed to transform a lemma to an inflected form. At prediction time, the authors employ a semi-Markov model (Sarawagi and Cohen, 2005) in order to choose an appropriate set of rules to be applied to produce a paradigm for a given lemma. The model was compared to Dreyer and Eisner (2011) on the CELEX dataset (Baayen et al., 1993) and shown to be superior. They were one of the first to use Wiktionary data in order to get inflectional paradigm tables and construct a language-independent model for morphological inflection. The model also outperforms a factored modification of the model where the rules are extracted separately on both CELEX and Wiktionary datasets.

Finally, some approaches such as Hulden et al. (2014) aim to capture variations within paradigms and then generalise them rather than learn a mapping between paradigms. Consider the following verbal inflections: `ring–rang–rung` and `sing–sang–sung`. We observe a similar form-changing pattern, i.e. (`x–`**`i`**`–y`, `x–`**`a`**`–y`, `x–`**`u`**`–y`), where `x` and `y` stand for variables that encode multiple possible forms. In order to capture such a pattern, the authors first find the longest common substring over multiple strings (which is NP-hard Maier (1978)) by intersecting FSAs that correspond to all substrings of all words. As a result, two patterns, `rng` and `swm` are extracted. Next, they fit them into a single table, replace each discontinuous sequence with a variable (e.g., `sw`→`x`, `m`→`y`) and, finally, merge the identical paradigms into a single one. Comparison with the approach of Durrett and DeNero (2013) shows that the current model performs slightly worse on a per-form prediction task while outperforms the earlier model on per-table predictions.

2.3.3.2 Large-scale Inflectional Paradigm Modeling

SIGMORPHON 2016 In the SIGMORPHON 2016 the participants were invited to submit systems in three different morphological reinflection tasks of increasing complexity. In the first task, the systems were provided with a lemma and the target tags, while in the second task, the lemma was replaced with another inflected form of the same lemma and its tags. In the last, the third task, the initial form tags were omitted, and the systems were only supplied with an initial inflected form and target tags. All three settings are illustrated in Table 2.11.

	Task 1	Task 2	Task3
Lemma	run	–	–
Source Tag	–	PAST	–
Source Form	–	*ran*	*ran*
Target Tag	PART.PRES	PART.PRES	PART.PRES
Target Form	*running*	*running*	*running*

Table 2.11 An example of three morphological reinflection tasks.

The systems produced either a single output string or a ranked list with maximum length of 20 predicted strings. In order to evaluate the performance, exact string accuracy (for the top predicted string), Levenshtein distance and reciprocal rank were measured.

The dataset comprised of 10 typologically diverse languages, including Arabic, Finnish, Georgian, Navajo, Maltese, Russian, Turkish, and Hungarian.

As a baseline the authors used a non-neural system, in particular, an FST with a perceptron classifier. Submitted systems can be classified into three groups. The first group, **align and transduce** (Alegria and Etxeberria, 2016; King, 2016; Liu and Mao, 2016; Nicolai et al., 2016), applied a pipeline approach. More specifically, they first trained an unsupervised alignment algorithm on the source-target pairs and extracted edit operations. Then they trained a discriminative model to apply the changes. Such an approach was initially inspired by Durrett and DeNero (2013) who first extracted a set of edit operations and then applied a semi-Markov CRF (Sarawagi and Cohen, 2005) to model the transformations. The transduction models were limited to the monotonic alignment case and were encoded by WFST (Mohri, 2002). The approaches from the second group, **RNN-based** (Aharoni et al., 2016; Kann and Schütze, 2016; Östling, 2016), were based on neural sequence-to-sequence models (Bahdanau et al., 2015; Sutskever et al., 2014) that have demonstrated great success in various NLP tasks in recent years (for instance, Faruqui et al. (2016) was one of the first to apply neural approaches to the morphological inflection task and show moderate success). The best performing system from Kann and Schütze (2016) was based on the sequence-to-sequence approach from Sutskever et al. (2014) with soft attention mechanism (Bahdanau et al., 2015) and GRUs (Cho et al., 2014) as basic units. The input form along with source and target tags was fed into the neural network as a single string (character-by-character),

and the network generated the target form on the character level as well. Essentially, for a paradigm with n elements the model reflects all possible n^2 mappings. Finally, some teams used **linguistically-inspired** approaches (Sorokin et al., 2017; Taji et al., 2016). The former segmented the forms into prefixes, stems and suffixes. A set of actions was then selected to be applied to stem to perform reinflection. The latter extracted a set of rules and then applied multi-way classification to select a set for a particular reinflection.

The neural approaches clearly outperformed non-neural models by a large margin (with a 13% gap between the best neural model and the best non-neural ones on average across all language and tasks). The latter ones generally used a pipeline, and the alignments were obtained independently of the transduction step, while neural systems jointly learned to align and transduce. Second, neural models allowed parameter sharing across different reinflections, which led to better generalisation. For the simple inflection task, the best system got 95% exact matches on average, ranging from 89% for Maltese to 99% for Hungarian. The results for ensemble, i.e. a setting when all systems' predictions are combined, demonstrate that non-neural approaches can still add some extra points to neural ones, and there is some room for improvement.

SIGMORPHON 2017. In the SIGMORPHON 2017 shared task (Cotterell et al., 2017a) the lemma form was always provided, and the systems were only required to perform morphological inflection or paradigm completion. For the inflection task, the individual forms were sparsely sampled from a large number of paradigms, and the systems did not necessarily observe any complete paradigms. For paradigm completion, the systems were trained on a few complete paradigms and filled gaps in the test paradigms. Monolingual unannotated data (the Wikipedia dump) was additionally provided.

The SIGMORPHON 2017 shared task was organised for 52 typologically different languages, including extremely low-resource such as Quechua, Haida, and Navajo.[28]

[28]Similar to the 2016 task, the data mainly comes from the Wiktionary. Data for Khaling, Kurmanji, Kurdish, Sorani Kurdish come from the Alexina project, `https://gforge.inria.fr/projects/alexina`, (Walther and Sagot, 2010; Walther et al., 2013), Haida was prepared by Jordan Lachler, and the Basque data was extracted from manually designed morphological FST (Alegria et al., 2009)

Due to the success of neural approaches in the 2016 task, most teams developed models based on neural architectures. Despite teams using similar architectures, the results varied substantially. Below, we discuss main the characteristics that led to the differences.

First of all, most systems implemented **neural parameterisation** and used some kind of a recurrent unit – either GRU (Chung et al., 2014), or LSTM (Hochreiter and Schmidhuber, 1997), and one team employed a convolutional neural network (CNN) (LeCun and Bengio, 1995). Generally, the neural networks had the target morphological tag and the source form as an input. The tag was usually segmented into subtags, and each subtag was assigned its own symbol.

Second, while most systems used **soft attention** mechanism (Bahdanau et al., 2015), few of them relied on modelling monotonic alignment with **hard attention**. This idea was introduced in the SIGMORPHON 2016 shared task by Aharoni et al. (2016). Indeed, the 2017 winning system used hard attention. Their system explicitly introduced a copy mechanism that substantially improved the performance in the low-resource setting.

Third, two systems used **reranking** of the output of a weaker system. One team employed a heuristically induced candidate set using the edit tree approach of Chrupała et al. (2008) and then chose the best edit tree, the other did linear reranking of top-k best output of the system.

The models were compared in high-(10,000 samples in the training data), medium-(1,000 samples), and low-resource (100 samples) settings. Neural systems excelled in the high resource setting. In the medium- and low-resource settings, standard encoder–decoder systems performed on par or even worse than the "align and transduce" baseline system of Liu and Mao (2016) if only training data is available. The systems that outperformed the baseline in these settings either successfully added some bias to their networks (such as a copy operation which is a common operation in morphological inflection), or synthesised extra data. The technique proved to be useful for languages with small and regular paradigms, while more complex paradigms, as in Latin, still require more variation in the training data.

Similarly to the SIGMORPHON 2016, an ensemble of all systems' predictions showed a substantial gain of accuracy, especially in medium (10%) and low (15%) settings, which

means that there was a lot of complementarity in their outputs, and possibly the generalisation patterns were different due to different "biases" the systems used.

The results reinforced the conclusion of the SIGMORPHON 2016 shared task that neural encoder–decoder models outperform other methods when a large amount of data is available. A bit surprisingly, the neural systems also performed reasonably well in low-resource settings under special training conditions mentioned above. The results obtained for the low-resource setting showed that for some large and less regular paradigms, a certain amount of variation in the training data is required. For instance, some cells of the paradigm are more informative.

SIGMORPHON 2018. The inflection subtask of SIGMORPHON 2018 shared task was identical to the previous year. The number of languages was extended and reached 103 (including extremely low-resource languages such as Murrinhpatha, Australian language). Similarly, the systems were evaluated in high-, medium-, and low-resource settings. Compared to 2017 shared task, results on 80% languages improved. For low-resource settings many systems either used additional artificial data, or attempted to learn sequences of edit operations to transform one form into another, or used pointer generator networks (See et al., 2017) which allow a copy mechanism. The second subtask focused on contextual reinflection and is discussed in Section 4.1 in more detail.

2.3.4 Derivational Models

Although in the last few years many neural morphological models have been proposed, most of them have focused on inflectional morphology (as shown in Section 2.3.3.2). Focusing on derivational processes, there are three main directions of research. The first deals with evaluation of word embeddings using a word analogy task (Gladkova et al., 2016). In this context, it has been shown that, unlike inflectional morphology, most derivational relations cannot be as easily captured using distributional methods. Researchers working on the second type of task attempt to predict derived forms using the embedding of its corresponding base form and a vector encoding a "derivational" shift. Guevara (2011) notes that derivational affixes can be modelled as a geometrical function over the vectors of the

base forms. On the other hand, Lazaridou et al. (2013) and Cotterell and Schütze (2018) represent derivational affixes as vectors and investigate various functions to combine them with base forms. Kisselew et al. (2015) and Padó et al. (2016) extend this line of research to model derivational morphology in German. This work demonstrates that various factors such as part of speech, semantic regularity and argument structure (Grimshaw, 1990) influence the predictability of a derived word. The third area of research focuses on the analysis of derivationally complex forms, which differs from this study in that we focus on generation. The goal of this line of work is to produce a canonicalised segmentation of an input word, e.g., `unhappiness`$\mapsto$`un-+happy+-ness` (Cotterell et al., 2015, 2016b). Note that the orthographic change `y`$\mapsto$`i` has been reversed.

In Lazaridou et al. (2013) the authors also point out the problem of word embeddings' quality which rapidly deteriorates with decrease of word frequency (as also shown in Bullinaria and Levy (2007)). In the paper, they specifically address a problem of sense prediction for morphological derivations, since it is quite common even in English (55% of the lemmata in CELEX database are morphologically complex, i.e are derived from other stems). They study various linear functions to estimate a derived word's vector such as multiplicative ($\mathbf{c} = \mathbf{uv}$), weighted additive ($\mathbf{c} = \alpha\mathbf{u} + \beta\mathbf{v}$ where α, β are some scalars), fully additive ($\mathbf{c} = A\mathbf{u} + B\mathbf{v}$ where A, B are weight matrices) and lexical function ($\mathbf{c} = U\mathbf{v}$ where U is a functor that corresponds to an affix). They evaluate base – derived word pairs for 18 English affixes extracted from CELEX dataset (Baayen et al., 1993). The results demonstrated that fully additive function leads to the best performance. Comparison of predictability of the final form's meaning depending on the affix attached showed that negation affixes (`in-`, `un-`, `-less`) are less compositional and more problematic, and on average they received lower base-relevance scores in human assessment. The highest scores were achieved by `-ness`, `-ity`, `-ist`, `-ion`, `-ness`, `-ful`, i.e ones that typically do not affect semantics but rather only change the part of speech.

The next section provides an overview of contemporary distributed (neural) models trained using both word- and subword-levels.

2.3.5 Distributed Representations and Distributional Semantics

2.3.5.1 Connectionism and Symbolism: On the Past Tense Debate

First, we start with some background and motivation for why we suggest connectionist approaches to be appropriate for modelling morphology. In the 1980s there was an active debate between connectionists and symbolists. Fodor and Pylyshyn (1988) provide a detailed comparison of two approaches to a cognitive architecture: connectionist and classical. Turing and von Neumann machines served as a basis for classical models of mind which operated on symbolic expressions, whereas connectionists' approach did not rely on storing, retrieving and operating on symbols but was rather focused on states. Unlike symbolists, connectionists studied states and, more essentially, what they represent. That is, they were (and are) following representational realism. But the key difference comes from what they assign semantic components to. Connectionists assign them to nodes, or groups of nodes, whereas classicists assign them to expressions, "i.e. to the sorts of things that get written on the tapes of Turing machines and stored at addresses in von Neumann machines" (Fodor and Pylyshyn, 1988, p. 12). Connectionists' models only account for causal connectedness that comes from activation propagation in the network. In addition to causal relations, classical models also account for structural relations.

A long discussion on connestionism versus symbolism (*"The past tense debate"*) started with Rumelhart and McClelland (1987) who proposed a neural network to simulate acquisition of the past tense inflection in English. Unlike more traditional models of that time, it did not rely on any symbolic rules, but rather was trained to mimic the rule-based behaviour. Based on these results, some researchers and philosophers (Bates and Elman, 1993; Churchland, 1996; Elman et al., 1998) argued that connectionism is an appropriate view on language knowledge representation and there is no need for explicit grammar. Elman (1993) also showed the models succeeded on subject-verb agreement in English. Their opponents, on the other hand, claimed that the connectionist models were incapable in realisation of some syntactic phenomena (Fodor and Pylyshyn, 1988; Fodor et al., 1974; Marcus, 2003). For instance, Marcus (2003) argued that the Elman's agreement model did not generalise to

the same extent as humans. A lot of criticism of Rumelhart and McClelland (1987)'s work in terms of their argumentation and conclusions has been expressed in Pinker and Prince (1988). For instance, they claim that the model does not learn many rules and is unable to explain differences between regular and irregular verbs as well as morphological and phonological regularities. In addition, it learns rules that do not exist in human language. Later, several models that addressed these issues were introduced (such as Plunkett and Marchman (1993) that imitated the U-shaped language acquisition in children). As Clahsen (1999) and Marcus (1998) note, the fundamental deficiency of these models is their reliance on particular statistical patterns in the data. Marcus additionally highlights that the models tend to overproduce English irregular patterns.

Nowadays, connectionsists' approaches have demonstrated a great success in many tasks, and in the following section we provide a description of some popular models that we will further use for our investigations.

2.3.5.2 Distributional Hypothesis

Many contemporary approaches in information retrieval, language modelling, or machine translation, are built on the **distributional hypothesis**, that states that a word and its context are similar, originates from Firth who stated that "a word is known by the company it keeps" (Firth, 1957, p. 11).

Harris formulated a clearer definition of distributional hypothesis by "All elements in a language can be grouped into classes whose relative occurrence can be stated exactly. However, for the occurrence of a particular member of one class relative to a particular member of another class, it would be necessary to speak in terms of probability, based on the frequency of that occurrence in a sample" (Harris, 1981, p. 146) and "The restrictions on relative occurrence of each element are described most simply by a network of interrelated statements, certain of them being put in terms of the results of certain others, rather than by a simple measure of the total restriction on each element separately" (Harris, 1981, p. 147).

Harrison has also added an additional hypothesis that co-occurrence of linguistic elements covers all the knowledge for a language without other types of information. Contrary to

that, psycholinguists generally differentiate two types of language, external, E-language, and internal, I-language (De Deyne et al., 2016). E-language refers to linguistic knowledge stored in datasets, corpora, i.e. externally. I-language corresponds to knowledge that is stored in the brain (which is often expressed by word association networks (De Deyne and Storms, 2008)). NLP models are E-language models in this sense, and in this thesis we only focus on E-language models.

2.3.5.3 Word-Level Models

Earlier approaches represented a word as a high-dimensional vector indicating whether or not it occurred in a context of other words of the vocabulary. The word co-occurrence matrix is defined as a Cartesian product over all words in the vocabulary. It consists of either a frequency of word collocations, or some form of pointwise mutual information. Since the matrix is large and sparse, various dimensionality reduction techniques (Singular Value Decomposition (Golub and Reinsch, 1970), Principal Component Analysis (Wold et al., 1987), Latent Semantic Analysis (Landauer et al., 1998)) were applied in order to reduce the size and identify the most important dimensions. During the last decade, the focus has changed to *learning* low-dimensional dense vectors, **word embeddings**, which we now review.

Curran (2004) focused on lexical semantics (synonymy extraction, in particular) and was one of the first to describe techniques of context learning and analyse how various types of extracted context affect word similarity. The proposed new context-weighted similarity metric significantly outperformed existing approaches.

Later Mnih and Hinton (2009) introduced `HLBL`, a log-bilinear formulation of an n-gram language model, which predicts the i-th word based on context words $(i-n,\ldots,i-2,i-1)$. This leads to the following training objective:

$$J = \frac{1}{T}\sum_{i=1}^{T}\frac{\exp(\tilde{\mathbf{w}}_i^\top \mathbf{w}_i + b_i)}{\sum_{k=1}^{V}\exp(\tilde{\mathbf{w}}_k^\top \mathbf{w}_k + b_k)},$$

where $\tilde{\mathbf{w}}_i = \sum_{j=1}^{n-1} C_j \mathbf{w}_{i-j}$ is the context embedding, C_j is a scaling matrix, and b_* is a bias term.

Collobert et al. (2011) proposed one of the first models, `SENNA`, a multi-task model for part of speech tagging, language modelling, chunking, and semantic role labelling. The model was one of the first to learn the features without any prior manual preprocessing. Its statistical language modelling component has a pairwise ranking objective to maximise the relative score of each word in its local context:

$$J = \frac{1}{T} \sum_{i=1}^{T} \sum_{k=1}^{V} \max \left[0, 1 - f(\mathbf{w}_{i-c}, \ldots, \mathbf{w}_{i-1}, \mathbf{w}_i) + f(\mathbf{w}_{i-c}, \ldots, \mathbf{w}_{i-1}, \mathbf{w}_k)\right],$$

where the last $c-1$ words are used as context, and $f(x)$ is a non-linear function of the input, defined as a multi-layer perceptron.

In 2013 Mikolov introduced another approach inspired by language modelling, `w2v` (Mikolov et al., 2013a,b), that predicts a word from its context (the CBoW model) with the objective:

$$J = \frac{1}{T} \sum_{i=1}^{T} \log \frac{\exp\left(\mathbf{w}_i^\top \sum\limits_{j \in [-c,+c], j \neq 0} \tilde{\mathbf{w}}_{i+j}\right)}{\sum_{k=1}^{V} \exp\left(\mathbf{w}_k^\top \sum\limits_{j \in [-c,+c], j \neq 0} \tilde{\mathbf{w}}_{i+j}\right)}$$

where $\mathbf{w}_i$ and $\tilde{\mathbf{w}}_i$ are the vector representations for the i-th word (as a focus or context word, respectively), V is the vocabulary size, T is the number of tokens in the corpus, and c is the context window size.[29] A similar model, skip-gram, predicted a context from a word.

Another successful model, `GloVe` (Pennington et al., 2014), is based on a similar bilinear formulation, framed as a low-rank decomposition of the matrix of corpus co-occurrence frequencies:

$$J = \frac{1}{2} \sum_{i,j=1}^{V} f(P_{ij})(\mathbf{w}_i^\top \tilde{\mathbf{w}}_j - \log P_{ij})^2,$$

where $\mathbf{w}_i$ is a vector for the left context, $\mathbf{w}_j$ is a vector for the right context, P_{ij} is the relative frequency of word j in the context of word i, and f is a heuristic weighting function to

[29] In a slight abuse of notation, the subscripts of $\mathbf{w}$ do double duty, denoting either the embedding for the i-th token, $\mathbf{w}_i$, or kth word type, $\mathbf{w}_k$.

balance the influence of high versus low term frequencies. Unlike `w2v`, it is not trained to predict words but rather learns to approximate the co-occurrence probability.

The `SVD` model (Levy et al., 2015a) uses positive pointwise mutual information (PMI) matrix defined as:

$$\text{PPMI}(w,c) = \max(\log \frac{\hat{P}(w,c)}{\hat{P}(w)\hat{P}(c)}, 0)\,,$$

where $\hat{P}(w,c)$ is the joint probability of word w and context c, and $\hat{P}(w)$ and $\hat{P}(c)$ are their marginal probabilities. The matrix is factorised by singular value decomposition.

Another recent generation of neural approaches mainly utilises various forms of Recurrent Neural Networks (RNNs) introduced in Elman (1990) and Jordan (1997). Luong et al. (2013); Mikolov et al. (2010); Socher et al. (2011, 2012, 2013b) are examples of this type. This stores the previous context and is essential for keeping track of the order of the words and potential modelling of **compositionality**.

2.3.5.4 Subword Representations

Most neural models for NLP rely on words as their basic units, and consequently face the problem of data sparsity, especially in morphologically rich languages. Both word co-occurrence matrices used in count-based approaches and lookup tables in neural ones face the problem of large size and inefficiency of memory usage. Typically, words are pruned by their frequency, so rare terms are often assigned a special `UNK` token, which comes at the expense of modelling accuracy. In order to address data sparsity and out-of-vocabulary (OOV) problems, words are often represented by sub-word units, typically morphemes, characters or ngrams.

Machine Translation and Language Modelling Decomposing a word into morphemes might seem to be the best and linguistically-inspired solution, but at the same time it requires a morphological analyser that might not be available for most languages. Therefore, often morphological analysers are replaced with automatic word segmentation tools such as Morfessor (Creutz and Lagus, 2007).

Luong et al. (2013) were one of the first to successfully tackle the problem of rare and unknown word representation in neural language models. They combined a recursive neural network which expressed a word's internal structure with a neural language model to represent sentence-level contextual information. Importantly, they showed that such a recursive architecture was appropriate for compositionality modelling. The idea behind their approach is quite straightforward, i.e. they estimate the probability of a word as some non-linearity f over a linear combination of its constituents: $f(W_m[\mathbf{x}_{stem};\mathbf{x}_{affix}]+b_m)$ where $W_m \in \mathbb{R}^{d\times 2d}$ corresponds to parameters of morpheme representations. The stem and affix vectors are applied recursively until the whole word embedding in eventually constructed. Importantly, the authors used automatic *word segmentations* obtained from Morfessor rather than linguistically motivated morphemes. Therefore, some of the segmentations were misleading, such as `de|fault|ed`. But still their models demonstrated a significant improvement over word-based ones on rare word similarity evaluation.

Botha and Blunsom (2014) modelled a complex word meaning as a linear combination of its constituents. The authors extended a log-bilinear model (Mnih and Hinton, 2007) to predict the next word as a linear function of its n preceding words, i.e. $\sum_{j=1}^{n-1}\mathbf{q}_j C_j$, where $C_j \in \mathbb{R}^{dxd}$ and $\mathbf{q}_j \in \mathbb{R}^d$ are context vectors. Each word vector they presented as a compositional function of its parts, namely a sum. They additionally added the word vector itself to account for non-compositional cases such as `greenhouse`. The word segmentation was automatically obtained from Morfessor (Creutz and Lagus, 2007). The authors also experimented on machine translation and showed consistent improvements in BLEU for English into Czech, Russian, German, Spanish and French.

Unsupervised analysers such as Morfessor are prone to segmentation errors, particularly on languages with fusional or non-concatenative morphology (e.g., templatic such as in Hebrew or Arabic). In these settings, character-level word representations may be more appropriate. One of the earliest solutions was presented by Schütze (1993) who proposed to replaced word co-occurrence with ngram (in particular, four-gram) co-occurrence matrices. The final representation of a word was calculated as a sum of representations of ngrams that

appeared within a context window for all occurrences of the word in corpus. The model was shown to achieve superior results on a word disambiguation task.

The transition from a word-level to character-level models was not straightforward. One of the first character-level statistical machine translation models (Vilar et al., 2007) failed to outperform word-based counterparts, and only a combined model outperformed the word-level in terms of BLEU score. Similarly, Neubig et al. (2013) were not able to show superior results to word-based models in a substring-level machine translation task.

Ling et al. (2015b) proposed the first successful character-level neural machine translation model. Their model comprises several RNNs with LSTM units. The source language context is obtained via bidirectional LSTMs (Hochreiter and Schmidhuber, 1997), consequently each word is represented by a linear combination of hidden states at a corresponding position. The target language context is represented as a forward LSTM (since the future context in yet to be predicted). The mapping between source and target representations is done by attentional mechanism as follows: $z_i = \mathbf{s}\tanh(W_t\mathbf{l}^f_{p-1} + W_s\mathbf{b}_i)$ and $a_i = \frac{\exp(z_i)}{\sum_{j\in[0,n]}\exp(z_j)}$ where $\mathbf{l}^f_{p-1}$ is a forward LSTM that encodes the target context translated up to $p-1$ position, $\mathbf{b}_i$ is a source vector, and $\mathbf{s}$ is a score vector. These attention weights are applied to the corresponding words in the source sentence and summed to get its representation. The novelty of the work comes from word representations. Instead of using a softmax over source and target vocabularies, the authors propose to compose the word representation from character-level LSTMs. More specifically, they obtain a word vector as a linear combination of the last hidden states of forward and backward LSTMs. On the target side, the word is generated character-by-character using the initial representation comprised of the average attention, current hidden state of the LSTM over the target sentence, and previously generated characters. The results obtained for English-French and English-Portuguese do not show a significant improvement over the word-based approach, but, as the authors mention, it is still much better than previous character-based models. Importantly, the authors show that the character-level model puts the orthographically similar words closer in space compared to word-based. The model also learns parts of inflectional morphology quite well (it correctly generates plural forms of nouns).

Other tasks also benefit from character-level representations. For instance, Lample et al. (2016) shows that incorporating biLSTM character-level word representations improves accuracy in a named entity recognition task. Several authors have proposed convolutional neural networks over character sequences as a part of models for part of speech tagging (Santos and Zadrozny, 2014), named entity recognition (Chiu and Nichols, 2016; Ma and Hovy, 2016), language (Kim et al., 2016) and machine translation (Belinkov et al., 2017; Costa-Jussà and Fonollosa, 2016). The latter presents an in-depth analysis of representations learned by neural MT models.

Sennrich et al. (2016) investigated various types of out-of-vocabulary words. Among them are named entities that should be either copied or transliterated, cognates and loanwords for which character-level translation rules might be sufficient, and, finally, morphologically complex words. In order to address such types of words, the authors proposed an efficient word segmentation method, *Byte Pair Encoding (BPE)*, originating from Gage (1994). The method evaluates character ngram frequencies, merges every frequent ngram, replacing it with a new symbol. They applied the method to segment a whole sentence into such pieces and evaluated it on English into German and Russian translation showing an improvement in terms of BLEU scores.

Bojanowski et al. (2017) introduced the FastText skip-gram model where they enriched the word2vec skip-gram model with subword representations obtained as a linear combination of constituent ngrams. Inspired by early non-neural n-gram based approaches such as Schütze (1993), the authors propose to instead a) extend it to a range from 1-grams to 6-grams; b) learn their neural representations. The model outperformed traditional CBoW and skip-gram models on word similarity tasks. The authors additionally demonstrated that the most important ngrams in a word correspond to morphemes. Grave et al. (2018) proposed a modification of the model, FastText CBoW, that uses the word2vec CBoW model with positional weights.

Finally, Vania and Lopez (2017) compared various types of word segmentations such as character n-grams, character-level models (CNN and BiLSTM), linguistically-inspired morphemes, and BPE and Morfessor segmentations on language modelling task. They conclude

that BiLSTMs composed of character trigrams outperform other character- and ngram-level models, but none of the models achieves high accuracy of linguistically-motivated morphological segmentation.

Many aforementioned models were shown to either perform similarly or even outperform standard word-level approaches. With a few notable exceptions (Heigold et al., 2017; Vania and Lopez, 2017; Yin et al., 2017), there was no systematic investigation of the various modelling architectures. In this thesis, we address the question of what linguistic aspects are best encoded in each type of architecture, and their efficacy as parts of a machine translation model when translating from morphologically rich languages.

POS tagging Finally, we would like to discuss subword-level approaches in POS tagging since the task is relevant to morphology learning. In POS tagging languages with largely fixed word order such as English typically rely on neighboring tags, but once we move to languages with flexible order, we would also need subword information to correctly guess a word's tag. Typically, in such languages word forms themselves are quite indicative of parts of speech. Santos and Zadrozny (2014) were one of the first to successfully apply character-level methods for POS tagging. In their approach, they used a concatenation of character-level CNN word representation and its representation as a whole unit. They followed Collobert et al. (2011) and used a context window approach assuming that a tag of a word mainly depends on its neighbors. Each sentence was eventually assigned with a tag path score, which has to be maximised. The authors showed that the model captured surface-level morphological patterns quite well. On the other hand, even though the majority of neighbours of `unsteadiness` followed the pattern `un-[stem]-ness`, some of them such as `business` were not related which raises a question of compositionality of word's meaning.

Ling et al. (2015a) proposed BiLSTM-based approaches for language modelling and POS tagging that further improved the results of Santos and Zadrozny (2014)'s CNN model as well as word-based ones and RNNs. They experimented with five Indo-European languages, namely English, Catalan, Portuguese, German, and Turkish. Their model considered compo-

sitional character-level word representation, i.e. it combined the last hidden states of forward and backward LSTMs running over word characters. The results showed that such a model obtained better perplexity on a language modelling task compared to a 5-gram Kneser-Ney and word-based model on all languages with significantly lower number of parameters.

Wang et al. (2015) proposed to use bidirectional LSTMs for sentential context representation achieving state-of-the-art accuracy for English (on WSJ data from Penn Treebank III (Marcus et al., 1993)) showing that without usage of any morphological features the system still can achieve good results.

Plank et al. (2016) compared a biLSTM tagger with CRF-based and TnT (Brants, 2000) taggers on 22 languages. The results show that word-level representations outperform character-level and combined word- and character-level representations. But on the other hand, the model that combines word, character and pre-trained embeddings performs the best and achieves higher accuracy compared to TnT and CRF. As expected, the most significant gains are among languages with complex morphology.

Moreover, very little has been studied in terms of fine-grained, or morphological, tagging evaluation. There, state-of-the-art results were achieved by a CRF tagger in Müller et al. (2013) and Müller and Schütze (2015). Heigold et al. (2017) were one of the first to perform a comparison of various character-level architectures on multiple (morphologically complex) languages for this task. First, they compared two types of character-level word representations: CNN-Highway (Kim et al., 2016) and forward LSTM. For sentential representations, they used a single biLSTM architecture. Eventually, they showed that the LSTM-based model consistently outperformed CNN-Highway as well as a non-neural CRF tagger, MarMot (Müller et al., 2013). Interestingly, the only exception was Arabic, possibly due to the templatic nature of morphology in Semitic languages. The authors also showed that the two architectures provide complementary information, and their combination leads to slight improvements of the accuracy.

Some work has been done on low-resource languages. Most low-resource languages belong to language families, and therefore could be linked to related languages via methods such as pre-training. Usually their morphology, word order and lexicon are quite similar

and therefore could be mapped into one another to some extent. How much improvement can we get this way? And how much data in low-resource languages do we need? These are questions addressed in Cotterell and Heigold (2017). In the paper, the authors jointly train the system for morphological prediction on the unified schema using the Universal Dependencies dataset (Nivre et al., 2016). Their results provide evidence for the hypothesis that relatedness plays a significant role and adds a strong extra signal for the morphological tagging task. Moreover, even in extremely low-resource settings, joint modelling improves accuracy in the case of closely related languages.

2.3.6 Learning of Compositionality

Finally, since we only deal with compositional morphology, we review compositionality more broadly and provide a list of approaches to model it. Compositionality is an essential concept for evaluation of a word, a sentence, a phrase, or a paragraph meaning. Its definition stands quite close to the distributional hypothesis. Frege's **Principle of Compositionality** states that the meaning of a complex item is a function of the meanings of its constituent expressions. Bach (1989) has added to the Principle an additional condition that the meaning also depends on the operations performed over those parts. Partee has put syntax in place of a set of operations, "The meaning of a complex expression is a function of the meaning of its parts and of the syntactic rules by which they are combined" (Partee, 1990, p. 318). Lakoff (1977) goes further and states that the whole meaning is greater than the parts' meanings.

While considering compositionality, it is a common and reasonable approach to assign an embedding vector of the same kind and dimension to the whole and its constituents. In relation to this, we should choose (1) how to collect words to be included into a composition, and (2) how the value of a composite vector is calculated. Note, that a large number of the approaches we have introduced are generally usable for decomposition of a word into morphemes (Botha and Blunsom, 2014) or even characters (Chen et al., 2015b; Ling et al., 2015a).

Let us first consider pairwise composition. As the procedure of composite embedding construction, one of the simplest approaches is merely **a weighted sum of elements**. Al-

though the Compositional Principle implies an operation to combine meanings, it tells us nothing about the operation itself; it remains arbitrary, and might differ radically from sense superposition. Therefore, many alternative formulas such as matrix multiplication and vector concatenation have been proposed. Smolensky (1990) proposed to use tensors, and Plate (1991) suggested holographic representations. Mitchell and Lapata (2010) presented a comprehensive study of various compositionality functions and showed superior results for weighted additive and multiplicative functions. It should be mentioned that one may attempt to discover non-compositional constructs like idioms by analysing the difference between the compositional embedding expected for a given context and one calculated from the component words with a trained composing function, similar to Salehi et al. (2014).

We have considered compositions of two components, but in real life we deal with longer sequences (like sentences). The most natural way for a sentence to be decomposed is building its parse tree and then following the structure of this tree while calculating compositional embeddings. In this approach, the resulting vector is produced by recursive computation of pairwise embeddings based on word embeddings as well as nested phrase embeddings calculated at previous steps. To learn parameters during such a variable-depth recursive procedure, one may employ Recursive Neural Networks (ReNNs).

It is common to rely on a parse tree provided by a separate pre-processing parsing stage. However, it may be worth combining parsing and embedding learning into a single loop, because not only does semantics depend on proper phrase detection, but also, vice versa, a phrase structure may depend on sense composition (Socher et al., 2011). An example of such an approach was proposed by Socher et al. (2010), Socher et al. (2011). In their algorithm, a composite embedding vector and a probability score of sharing the same phrase are both evaluated for each pair of adjacent words using a single neural network layer. Words and phrases are connected into parent phrases recursively using either the right or left neighbour according to their probability scores in order to build a predicted parse tree.

There are also approaches that are generally similar but rely less on parse trees, including ones that group words merely based on POS tagging and/or WordNet clustering instead of full parsing (Chen et al., 2015a). Word clustering also may be used in the compositional function

in order to improve the performance, as a word-based approach requires normalisation against a sum over the whole lexicon.

Parse tree based approaches may capture long-range dependencies, but their complexity is high. There are also many techniques which do not rely on a parse tree at all and, in such cases, are free of recursive branching, thus being faster and allowing a variety of common machine learning algorithms to be used. A basic approach of such kind uses a contextual window of a fixed length to select words so that all content fitting to the window around a given position in a text is considered as a composition, with (possibly) distinct positions for every component word. This technique yields fairly precise results at low cost but lacks many aspects dependent on surrounding words (phrase structure, discourse, etc.). In order to include more surrounding words into the scope of computation while avoiding recursive variable-size frames, extra words beyond the context window may be treated as a bag-of-words (Huang et al., 2012). In this case, a detailed consideration of parse tree structure with respect to a given text position may be replaced by a combination of two context flavours, the former being a fixed length sequence of words representing the local phrase structure while the latter being a much larger bag-of-words representing the global context. This behaves like a combination of a context-free grammar and RNN.[30]

Finally, Lake and Baroni (2018) systematically compared a number of RNNs on a new compositionality modelling task. They presented a toy dataset consisting of samples such as "jump $\rightarrow$ JUMP", "turn left twice $\rightarrow$ LTURN LTURN", "jump thrice $\rightarrow$ JUMP JUMP JUMP". The objective is to translate the commands on the left to the set of actions on the right. The task measures both generalisability and awareness of compositionality of the models. In their experiments, the authors studied a number of RNN-based models such as simple RNNs (Elman, 1990), LSTMs (Hochreiter and Schmidhuber, 1997), GRUs (Chung et al., 2014), each with and without attention (Bahdanau et al., 2015). LSTMs without attention showed overall the best results in a series of experiments. They find that the models are able to

[30]Note that the consideration of context window content is not a trivial task, and there are different approaches to calculate its composition. It has been proposed to multiply a matrix (an "adjacency matrix" that needs to be learnt) by the concatenation of word embedding vectors; in general, the problem is similar to one considered above for two-component composition vector construction (however, learning is an extra challenge as it deals with arbitrary window context, not necessarily with true word composition).

do zero-shot generalisations when commands from training and test sets are similar. More importantly, they conduct an experiment on generalisation of composition across primitive actions, i.e. they train the models on commands such as “run”, “run twice”, “jump” and then require the models to predict all presented compositions on the action that they only observed in the primitive context (e.g., produce “jump twice”). They find that in this task RNNs fail. They conclude that current models lack the ability to extract systematic rules from the data, i.e. the ability to see a pattern such as “translate(x and y)=translate(x)translate(y)”.

2.4 Conclusion

In this chapter, we introduced two types of morphology, inflectional and derivational. We discussed their similarities and differences, and several linguistic theories of treating morphology: Item and Process and Item and Arrangement. We introduced the datasets and tasks used in prior work and that we will be evaluating against and targeting in our experiments. In the second part of the chapter, we introduced main tools and approaches used in NLP for morphology modelling such as FSTs, rule-based and neural models, and provided a comparison of them on morphological reinflection task studied in terms of SIGMORPHON shared tasks. The final part of the chapter provided a summary of distributed models that are based on distributional semantics principle and discussed approaches to compositionality modelling.

Chapter 3

Evaluation of Word Embeddings and Distributional Semantics Approach

A large part of the chapter appears in the following papers:

Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1671–1682, 2016.

Ekaterina Vylomova, Trevor Cohn, Xuanli He, and Gholamreza Haffari. Word representation models for morphologically rich languages in neural machine translation. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 103–108, 2017a.

3.1 Introduction

In this chapter, we evaluate word- and character-level word embeddings. In the first part of the chapter, we provide a comparative analysis of the word embeddings from a number of models. We evaluate their performance at capturing lexical semantic (hypernymy,

meronymy), morphosemantic (verb-to-noun nominalisation), and morphosyntactic (transformation of a verb from present tense to past tense) relations. We only focus on asymmetric binary relations that are expressed as the difference between corresponding word vectors which, as has been shown, is one of the best functions to model these relations. Our results show that word-level models perform well at capturing morphosyntactic relations, while lexical semantic and morphosemantic ones are more challenging in terms of generalisation and require more specific training conditions. We also demonstrate a large gap between the models' performance in the setting when *each* word pair corresponds to a relation from the initial set and a more realistic situation when the models are also provided with many non-existent relations. In the latter one, the models perform poorly, and we suggest usage of negative sampling technique similar to the one used for training of the models in order to improve their generalisation ability.

In our comparison of word- and character-level models we show that character-level models typically perform better, and the character-level models (`fasttext`, in particular) trained on a smaller size corpus perform nearly as well as the word-level trained on a much larger one.

Finally, we translate the initial dataset into Russian and compare the performance of character-level models. Our observations are similar to those for English, although for morphosyntatic relations the accuracy is quite close to 100%, which we ascribe to the fact that Russian word forms are less ambiguous and more indicative of morphosyntactic properties.

In the second part of the chapter, we evaluate word- and character-level models on a machine translation task. We enrich word-level representations on the encoder part with two types of character-level representations, CNN and biLSTM. We experiment on translation from morphological rich languages (Russian and Estonian) into English and show that character-level models improve the word representation for low-frequency tokens. We show that CNN representations are more focused on the lemma (central) part, while biLSTM-based representations are focused on the periphery of the word. We additionally demonstrate that the biLSTM model is superior to the CNN for these languages in terms of capturing

morphological similarities (which could be explained by the fact that they are both based on concatenative word formation). In addition, we initialise the source word embeddings with the pre-trained character-level ones used for evaluation in the previous part. We show that the models achieve results close to the biLSTM model.

3.2 Language Modelling for English

Learning to identify lexical relations is a fundamental task in NLP, and can contribute to many NLP applications including paraphrasing and generation, machine translation, and ontology building (Banko et al., 2007; Hendrickx et al., 2010).

Recently, as discussed in Section 2.3.5, attention has been focused on identifying lexical relations using word embeddings, which are dense, low-dimensional vectors obtained either from a "predict-based" neural network trained to predict word contexts, or a "count-based" traditional distributional similarity method combined with dimensionality reduction. The skip-gram model of Mikolov et al. (2013a) and other similar language models have been shown to perform well on an analogy completion task (Levy and Goldberg, 2014a; Mikolov et al., 2013b,c), in the context of *relational similarity* prediction (Turney, 2006), where the task is to predict the missing word in analogies such as $A{:}B :: C{:}{-}?{-}$. A well-known example involves predicting the vector **queen** from the vector combination $\mathbf{king} - \mathbf{man} + \mathbf{woman}$, where linear operations on word vectors appear to capture the lexical relation governing the analogy, in this case OPPOSITE-GENDER. The results extend to several semantic relations such as CAPITAL-OF ($\mathbf{paris} - \mathbf{france} + \mathbf{poland} \approx \mathbf{warsaw}$)[1] and morphosyntactic relations such as PLURALISATION ($\mathbf{cars} - \mathbf{car} + \mathbf{apple} \approx \mathbf{apples}$). Remarkably, since the model is not trained for this task, the relational structure of the vector space appears to be an emergent property.

The key operation in these models is *vector difference*, or *vector offset*. For example, the $\mathbf{paris} - \mathbf{france}$ vector appears to encode CAPITAL-OF, presumably by cancelling out the features of **paris** that are France-specific, and retaining the features that distinguish a

[1] Case-folding is applied during pre-processing step.

capital city (Levy and Goldberg, 2014a). The success of the simple offset method on analogy completion suggests that the difference vectors ("DIFFVEC" hereafter) must themselves be meaningful: their direction and/or magnitude encodes a lexical relation.

Previous analogy completion tasks used with word embeddings have limited coverage of lexical relation types. Moreover, the task does not explore the full implications of DIFFVECs as meaningful vector space objects in their own right, because it only looks for a one-best answer to the particular lexical analogies in the test set. In this chapter, we introduce a new, larger dataset covering many well-known lexical relation types from the linguistics and cognitive science literature. We then apply DIFFVECs to two new tasks: unsupervised and supervised relation extraction. First, we cluster the DIFFVECs to test whether the clusters map onto true lexical relations. We find that the clustering works remarkably well, although syntactic relations are captured better than semantic ones.

Second, we perform classification over the DIFFVECs and obtain remarkably high accuracy in a closed-world setting (over a predefined set of word pairs, each of which corresponds to a lexical relation in the training data). When we move to an open-world setting including random word pairs — many of which do not correspond to any lexical relation in the training data — the results are poor. We then investigate methods for better attuning the learned class representation to the lexical relations, focusing on methods for automatically synthesising negative instances. We find that this improves the model performance substantially.

We also find that hyper-parameter optimised count-based methods are competitive with predict-based methods under both clustering and supervised relation classification, in line with the findings of Levy et al. (2015a).

3.2.1 Relation Learning

A lexical relation is a binary relation r holding between a word pair (w_i, w_j); for example, the pair $(cart, wheel)$ stands in the WHOLE-PART relation. Relation learning in NLP includes relation extraction, relation classification, and relational similarity prediction. In relation extraction, related word pairs in a corpus and the relevant relation are identified. Given a word pair, the relation classification task involves assigning a word pair to the correct relation

from a pre-defined set. In the Open Information Extraction paradigm (Banko et al., 2007; Weikum and Theobald, 2010), also known as unsupervised relation extraction, the relations themselves are also learned from the text (e.g. in the form of text labels). On the other hand, relational similarity prediction involves assessing the degree to which a word pair (A,B) stands in the same relation as another pair (C,D), or to complete an analogy $A{:}B :: C{:}{-?-}$. Relation learning is an important and long-standing task in NLP and has been the focus of a number of shared tasks (Girju et al., 2007; Hendrickx et al., 2010; Jurgens et al., 2012).

Recently, attention has turned to using vector space models of words for relation classification and relational similarity prediction. Distributional word vectors have been used for detection of relations such as hypernymy (Geffet and Dagan, 2005; Kotlerman et al., 2010; Lenci and Benotto, 2012; Rimell, 2014; Santus et al., 2014; Weeds et al., 2014) and qualia structure (Yamada et al., 2009). An exciting development has been the demonstration that vector difference over word embeddings (Mikolov et al., 2013c) can be used to model word analogy tasks. This has given rise to a series of papers exploring the DIFFVEC idea in different contexts. The original analogy dataset has been used to evaluate predict-based language models by Mnih and Kavukcuoglu (2013) and also Zhila et al. (2013), who combine a neural language model with a pattern-based classifier. Kim and de Marneffe (2013) use word embeddings to derive representations of adjective scales, e.g. *hot—warm—cool—cold*. Fu et al. (2014) similarly use embeddings to predict hypernym relations, in this case clustering words by topic to show that hypernym DIFFVECs can be broken down into more fine-grained relations. Neural networks have also been developed for joint learning of lexical and relational similarity, making use of the WordNet relation hierarchy (Bordes et al., 2013; Faruqui et al., 2015; Fried and Duh, 2015; Socher et al., 2013a; Xu et al., 2014; Yu and Dredze, 2014).

Another strand of work corresponding to the vector difference approach has analysed the structure of predict-based embedding models in order to help explain their success on the analogy and other tasks (Arora et al., 2015; Levy and Goldberg, 2014a,b). However, there has been no systematic investigation of the range of relations for which the vector difference method is most effective, although there have been some smaller-scale investigations in this

direction. Makrai et al. (2013) divide antonym pairs into semantic classes such as quality, time, gender, and distance, finding that for about two-thirds of antonym classes, DIFFVECs are significantly more correlated than random. Necşulescu et al. (2015) train a classifier on word pairs, using word embeddings to predict coordinates, hypernyms, and meronyms. Roller and Erk (2016) analyse the performance of vector concatenation and difference on the task of predicting lexical entailment and show that vector concatenation overwhelmingly learns to detect Hearst patterns (e.g., *including*, *such as*). Köper et al. (2015) undertake a systematic study of morphosyntactic and semantic relations on word embeddings produced with `word2vec` ("`w2v`" hereafter; see Section 3.2.2.1) for English and German. They test a variety of relations including word similarity, antonyms, synonyms, hypernyms, and meronyms, in a novel analogy task. Although the set of relations tested by Köper et al. (2015) is somewhat more constrained than the set we use, there is a good deal of overlap. However, their evaluation is performed in the context of relational similarity, and they do not perform clustering or classification on the DIFFVECs.

Recently, Hakami et al. (2018) published an extensive study of DIFFVECs and their mathematical basis. Here, we also provide our analysis why vector differences seem to be a promising approach for binary relation classification of word vectors obtained by training language models.

In a traditional language model, we estimate the probability of the next word given its prior context, i.e. $p(w_i = w|w_{1:i-1}) = \frac{\exp(\mathbf{w}^\top \mathbf{c})}{\sum_{w' \in V} exp(\mathbf{w'}^\top \mathbf{c})}$, where V is a size of the lexicon and $\mathbf{c}$ is a vector that represents a context. It could be a weighted sum of k previous words as in a log-bilinear model, or a vector corresponding to the network's hidden state (that, in turn, depends on the weighted sum of its current input and the previous state) with some weight as in RNNs. Contrary to that, `w2v` model predicts the *center* word from its surrounding context of some fixed size. The context is expressed as a sum of word vectors. In general, we see language modelling as a case of multinomial logistic regression over V possible classes. Paraphrasing the task, we aim to solve $V-1$ binary regression tasks each of which could be expressed as follows: $\log \frac{p(w_i=w|c)}{p(w_i=w'|c)} = \mathbf{c} \cdot \theta_{\mathbf{w}}$, where $\mathbf{c}$ is an input vector (in LMs it is usually

context representation), $\theta \in \mathbb{R}^{D \times V}$ are the model's parameters and $\theta_{\mathbf{w}}$ corresponds to the parameters of the class w.

For the `w2v` CBoW we can rewrite it as follows:

$$\log \frac{p(w_i = w|c)}{p(w_i = w'|c)} = \log \frac{\exp\left(\mathbf{w}^\top \sum\limits_{j \in [-c,+c], j \neq 0} \tilde{\mathbf{w}}_{i+j}\right)}{\exp\left(\mathbf{w}'^\top \sum\limits_{j \in [-c,+c], j \neq 0} \tilde{\mathbf{w}}_{i+j}\right)} =$$

$$= \left(\mathbf{w}^\top \sum_{j \in [-c,+c], j \neq 0} \tilde{\mathbf{w}}_{i+j}\right) - \left(\mathbf{w}'^\top \sum_{j \in [-c,+c], j \neq 0} \tilde{\mathbf{w}}_{i+j}\right)$$

And, finally, we rewrite it as:

$$\log \frac{p(w_i = w|c)}{p(w_i = w'|c)} = \sum_{j \in [-c,+c], j \neq 0} \tilde{\mathbf{w}}_{i+j} \left(\mathbf{w}^\top - \mathbf{w}'^\top\right)$$

In the case of an RNN it changes to the following:

$$\log \frac{p(w_i = w|c)}{p(w_i = w'|c)} = (\beta_{\mathbf{w}} \cdot \mathbf{h_t} + \mathbf{b_w}) - (\beta_{\mathbf{w}'} \cdot \mathbf{h_t} + \mathbf{b_{w'}})$$

As we see, in both cases log-odds ratio can be re-written as a vector difference.

3.2.2 General Approach and Resources

We define the task of lexical relation learning to take a set of (ordered) word pairs $\{(w_i, w_j)\}$ and a set of binary lexical relations $R = \{r_k\}$, and map each word pair (w_i, w_j) as follows: (a) $(w_i, w_j) \mapsto r_k \in R$, i.e. the "closed-world" setting, where we assume that all word pairs can be uniquely classified according to a relation in R; or (b) $(w_i, w_j) \mapsto r_k \in R \cup \{\phi\}$ where ϕ signifies the fact that none of the relations in R apply to the word pair in question, i.e. the "open-world" setting.

Our starting point for lexical relation learning is the assumption that important information about various types of relations is implicitly embedded in the offset vectors. While a range of

methods have been proposed for composing word vectors (Baroni et al., 2012; Roller et al., 2014; Weeds et al., 2014), in this research we focus exclusively on DIFFVEC (i.e. $\mathbf{w_2} - \mathbf{w_1}$). A second assumption is that there exist dimensions, or directions, in the embedding vector spaces responsible for a particular lexical relation. Such dimensions could be identified and exploited as part of a clustering or classification method, in the context of identifying relations between word pairs or classes of DIFFVECs.

In order to test the generalisability of the DIFFVEC method, we require: (1) word embeddings, and (2) a set of lexical relations to evaluate against. As the focus of this chapter is not the word embedding pre-training approaches so much as the utility of the DIFFVECs for lexical relation learning, we take a selection of several pre-trained word embeddings with strong currency in the literature, as detailed in Section 2.3.5. We also include the state-of-the-art count-based approach of Levy et al. (2015a), to test the generalisability of DIFFVECs to count-based word embeddings.

For the lexical relations, we want a range of relations that is representative of the types of relational learning tasks targeted in the literature, and where there is availability of annotated data. To this end, we construct a dataset from a variety of sources, focusing on lexical semantic relations (which are less well represented in the analogy dataset of Mikolov et al. (2013c)), but also including morphosyntactic and morphosemantic relations (see Section 3.2.2.1).

3.2.2.1 Word Embeddings

We consider four highly successful word embedding models in our experiments: `w2v` (Mikolov et al., 2013a,b), `GloVe` (Pennington et al., 2014), `SENNA` (Collobert and Weston, 2008), and `HLBL` (Mnih and Hinton, 2009). We also include `SVD` (Levy et al., 2015a), a count-based model which factorises a positive PMI (PPMI) matrix. For consistency of comparison, we train `SVD` as well as a version of `w2v` and `GloVe` (which we call $\texttt{w2v}_{\text{wiki}}$ and $\texttt{GloVe}_{\text{wiki}}$, respectively) on a fixed English Wikipedia corpus (comparable in size to the training data of `SENNA` and `HLBL`), and apply the preprocessing of Levy et al. (2015a). We

Name	Dimensions	Training data
w2v	300	100×10^9
GloVe	200	6×10^9
SENNA	100	37×10^6
HLBL	200	37×10^6
FT_{sg}	300	50×10^6
FT_{cbow}	300	N/A
w2v_{wiki}	300	50×10^6
$\text{GloVe}_{\text{wiki}}$	300	50×10^6
SVD_{wiki}	300	50×10^6

Table 3.1 The pre-trained word embeddings used in our experiments, with the number of dimensions and size of the training data (in word tokens). The models trained on English Wikipedia ("wiki") are in the lower half of the table.

additionally normalise the w2v_{wiki} and SVD_{wiki} vectors to unit length; $\text{GloVe}_{\text{wiki}}$ is natively normalised by column.[2]

For HLBL and SENNA, we use the pre-trained embeddings from Turian et al. (2010), trained on the Reuters English newswire corpus. In both cases, the embeddings were scaled by the global standard deviation over the word-embedding matrix, $W_{\text{scaled}} = 0.1 \times \frac{W}{\sigma(W)}$.

For character-level models we use FastText embeddings, which are a modification of w2v. In particular, we compare two pre-trained models, FT_{sg} (Bojanowski et al., 2017) and FT_{cbow} (Grave et al., 2018), corresponding to skip-gram and positional CBoW, respectively.

For w2v_{wiki}, $\text{GloVe}_{\text{wiki}}$ and SVD_{wiki} we used English Wikipedia. We followed the same preprocessing procedure described in Levy et al. (2015a),[3] i.e., lower-cased all words and removed non-textual elements. During the training phase, for each model we set a word frequency threshold of 5. For the SVD model, we followed the recommendations of Levy et al. (2015a) in setting the context window size to 2, negative sampling parameter to 1, eigenvalue weighting to 0.5, and context distribution smoothing to 0.75; other parameters were assigned their default values. For the other models we used the following parameter

[2] We ran a series of experiments on normalised and unnormalised w2v models, and found that normalisation tends to boost results over most of our relations (with the exception of $\text{LexSem}_{\text{Event}}$ and $\text{Noun}_{\text{Coll}}$). We leave a more detailed investigation of normalisation to future work.

[3] Although the w2v model trained without preprocessing performed marginally better, we used preprocessing throughout for consistency.

Relation	Description	Example
LEXSEM$_{\text{Hyper}}$	hypernym	(*animal*, *dog*)
LEXSEM$_{\text{Mero}}$	meronym	(*airplane*, *cockpit*)
LEXSEM$_{\text{Attr}}$	characteristic quality, action	(*cloud*, *rain*)
LEXSEM$_{\text{Cause}}$	cause, purpose, or goal	(*cook*, *eat*)
LEXSEM$_{\text{Space}}$	location or time association	(*aquarium*, *fish*)
LEXSEM$_{\text{Ref}}$	expression or representation	(*song*, *emotion*)
LEXSEM$_{\text{Event}}$	object's action	(*zip*, *coat*)
NOUN$_{\text{SP}}$	plural form of a noun	(*year*, *years*)
VERB$_{3}$	V;1SG;PRES → V;3SG;PRES	(*run*, *runs*)
VERB$_{\text{Past}}$	V;1SG;PRES →V;PAST	(*run*, *ran*)
VERB$_{\text{3Past}}$	V;3SG;Pres → V;PAST	(*runs*, *ran*)
LVC	light verb construction	(*give*, *approval*)
VERBNOUN	nominalisation of a verb	(*approve*, *approval*)
PREFIX	prefixing with *re* morpheme	(*vote*, *revote*)
NOUN$_{\text{Coll}}$	collective noun	(*army*, *ants*)

Table 3.2 Description of the 15 lexical relations (see continuation on Table 3.3).

values: for `w2v`, context window = 8, negative samples = 25, hs = 0, sample = 1e-4, and iterations = 15; and for `GloVe`, context window = 15, x_max = 10, and iterations = 15.

Lexical Relations In order to evaluate the applicability of the DIFFVEC approach to relations of different types, we assembled a set of lexical relations in three broad categories: lexical semantic relations, morphosyntactic paradigm relations, and morphosemantic relations. We constrained the relations to be binary and to have fixed directionality.[4] Consequently we excluded symmetric lexical relations such as synonymy. We additionally constrained the dataset to the words occurring in all embedding sets. There is some overlap between our relations and those included in the analogy task of Mikolov et al. (2013c), but we include a much wider range of lexical semantic relations, especially those standardly evaluated in the relation classification literature. We manually filtered the data to remove duplicates (e.g., as part of merging the two sources of LEXSEM$_{\text{Hyper}}$ intances), and normalise directionality.

[4]Word similarity is not included; it is not easily captured by DIFFVEC since there is no homogeneous "content" to the lexical relation which could be captured by the direction and magnitude of a difference vector (other than that it should be small).

Relation	Pairs	Source
LEXSEM$_{\text{Hyper}}$	1173	SemEval'12 + BLESS
LEXSEM$_{\text{Mero}}$	2825	SemEval'12 + BLESS
LEXSEM$_{\text{Attr}}$	71	SemEval'12
LEXSEM$_{\text{Cause}}$	249	SemEval'12
LEXSEM$_{\text{Space}}$	235	SemEval'12
LEXSEM$_{\text{Ref}}$	187	SemEval'12
LEXSEM$_{\text{Event}}$	3583	BLESS
NOUN$_{\text{SP}}$	100	MSR
VERB$_{3}$	99	MSR
VERB$_{\text{Past}}$	100	MSR
VERB$_{\text{3Past}}$	100	MSR
LVC	58	Tan et al. (2006b)
VERBNOUN	3303	WordNet
PREFIX	118	Wiktionary
NOUN$_{\text{Coll}}$	257	Web source

Table 3.3 Number of samples and sources of the 15 lexical relations.

The final dataset consists of 12,458 triples $\langle \text{relation}, \text{word}_1, \text{word}_2 \rangle$, comprising 15 relation types, extracted from SemEval'12 (Jurgens et al., 2012), BLESS (Baroni and Lenci, 2011), the MSR analogy dataset (Mikolov et al., 2013c), the light verb dataset of Tan et al. (2006a), Princeton WordNet (Fellbaum, 1998), Wiktionary,[5] and a web lexicon of collective nouns,[6] as listed in Tables 3.2 and 3.3.[7]

3.2.3 Clustering

Assuming DIFFVECs are capable of capturing all lexical relations equally, we would expect clustering to be able to identify sets of word pairs with high relational similarity, or equivalently clusters of similar offset vectors. Under the additional assumption that a given word pair corresponds to a unique lexical relation (in line with our definition of the lexical relation learning task in Section 3.2.2), a hard clustering approach is appropriate. In order to test these assumptions, we cluster our 15-relation closed-world dataset in the first instance, and evaluate against the lexical resources in Section 3.2.2.1.

[5] `http://en.wiktionary.org`

[6] `http://www.rinkworks.com/words/collective.shtml`

[7] The dataset is available at `http://github.com/ivri/DiffVec`

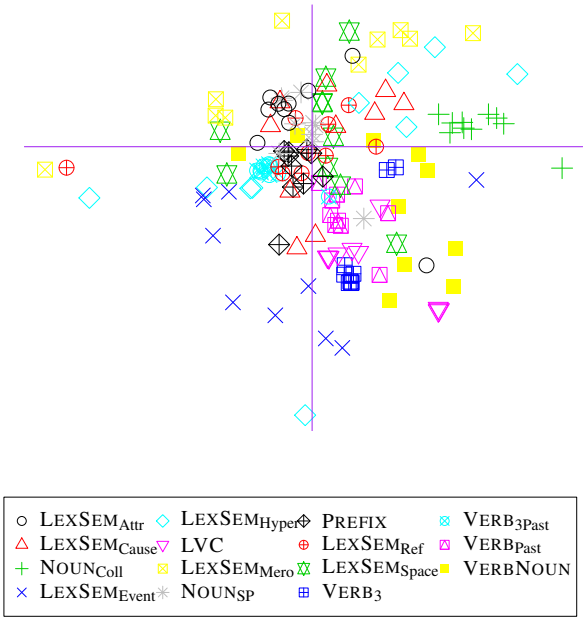


Figure 3.1 t-SNE projection (Van der Maaten and Hinton, 2008) of DIFFVECs for 10 sample word pairs of each relation type, based on `w2v`. The intersection of the two axes identify the projection of the zero vector. Best viewed in colour.

As further motivation, we projected the DIFFVEC space for a small number of samples of each class using t-SNE (Van der Maaten and Hinton, 2008), and found that many of the morphosyntactic relations (VERB$_3$, VERB$_{Past}$, VERB$_{3Past}$, NOUN$_{SP}$) form tight clusters (Figure 3.1).

We cluster the DIFFVECs between all word pairs in our dataset using spectral clustering Von Luxburg (2007). Spectral clustering has two hyperparameters: the number of clusters, and the pairwise similarity measure for comparing DIFFVECs. We tune the hyperparameters over development data, in the form of 15% of the data obtained by random sampling, selecting the configuration that maximises the V-Measure (Rosenberg and Hirschberg, 2007). Figure 3.2 presents V-Measure values over the test data for each of the four word embedding models. We show results for different numbers of clusters, from $N = 10$ in steps of 10, up to $N = 80$ (beyond which the clustering quality diminishes).[8] Observe that `w2v` and `fasttext` achieve the best results, with the highest V-Measure value for `w2v` of around

[8] Although 80 clusters $\gg$ our 15 relation types, the SemEval'12 classes each contain numerous subclasses, so the larger number may be more realistic.

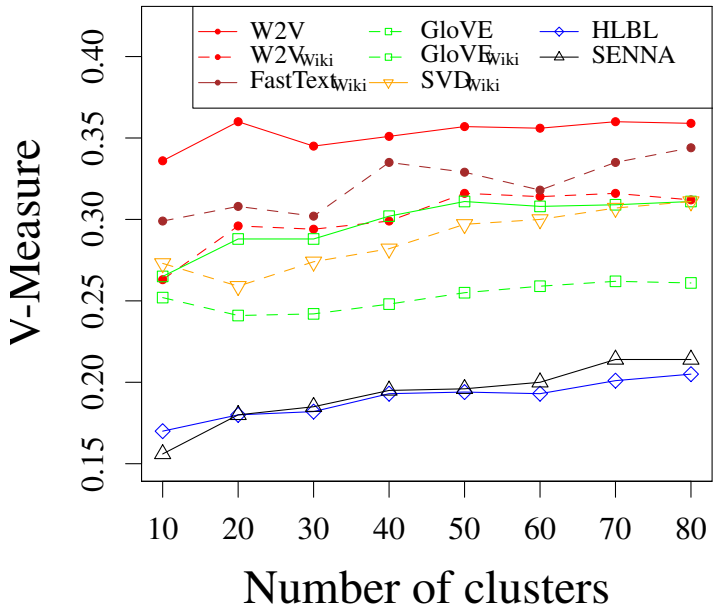


Figure 3.2 Spectral clustering results, comparing cluster quality (V-Measure) and the number of clusters. DIFFVECs are clustered and compared to the known relation types. Each line shows a different source of word embeddings.

0.36,[9] which is relatively constant over varying numbers of clusters. `GloVe` and `SVD` mirror this result, but are consistently below `w2v` at a V-Measure of around 0.31. `HLBL` and `SENNA` performed very similarly, at a substantially lower V-Measure than `w2v` or `GloVe`, closer to 0.21. As a crude calibration for these results, over the related clustering task of word sense induction, the best-performing systems in SemEval-2010 Task 4 (Manandhar et al., 2010) achieved a V-Measure of under 0.2.

The lower V-measure for $\texttt{w2v}_{\text{wiki}}$ and $\texttt{GloVe}_{\text{wiki}}$ (as compared to `w2v` and `GloVe`, respectively) indicates that the volume of training data plays a role in the clustering results. However, both methods still perform well above `SENNA` and `HLBL`, and `w2v` has a clear empirical advantage over `GloVe`. We note that $\texttt{SVD}_{\text{wiki}}$ performs almost as well as $\texttt{w2v}_{\text{wiki}}$, consistent with the results of Levy et al. (2015a).

We additionally calculated the entropy for each lexical relation, based on the distribution of instances belonging to a given relation across the different clusters (and simple MLE). For

[9] V-Measure returns a value in the range $[0, 1]$, with 1 indicating perfect homogeneity and completeness.

	w2v	GloVe	HLBL	SENNA
LEXSEM$_{Attr}$	**0.49**	0.54	0.62	0.63
LEXSEM$_{Cause}$	**0.47**	0.53	0.56	0.57
LEXSEM$_{Space}$	**0.49**	0.55	0.54	0.58
LEXSEM$_{Ref}$	**0.44**	0.50	0.54	0.56
LEXSEM$_{Hyper}$	0.44	0.50	**0.43**	0.45
LEXSEM$_{Event}$	**0.46**	0.47	0.47	0.48
LEXSEM$_{Mero}$	**0.40**	0.42	0.42	0.43
NOUN$_{SP}$	**0.07**	0.14	0.22	0.29
VERB$_3$	**0.05**	0.06	0.49	0.44
VERB$_{Past}$	**0.09**	0.14	0.38	0.35
VERB$_{3Past}$	0.07	**0.05**	0.49	0.52
LVC	**0.28**	0.55	0.32	0.30
VERBNOUN	**0.31**	0.33	0.35	0.36
PREFIX	0.32	**0.30**	0.55	0.58
NOUN$_{Coll}$	**0.21**	0.27	0.46	0.44

Table 3.4 The entropy for each lexical relation over the clustering output for each set of pre-trained word embeddings.

each embedding method, we present the entropy for the cluster size where V-measure was maximised over the development data. Since the samples are distributed nonuniformly, we normalise entropy results for each method by $\log(n)$ where n is the number of samples in a particular relation. The results are in Table 3.4, with the lowest entropy (purest clustering) for each relation indicated in bold.

Looking across the different lexical relation types, the morphosyntactic paradigm relations (NOUN$_{SP}$ and the three VERB relations) are by far the easiest to capture. The lexical semantic relations, on the other hand, are the hardest to capture for all embeddings.

Considering `w2v` embeddings, for VERB$_3$ there was a single cluster consisting of around 90% of VERB$_3$ word pairs. Most errors resulted from POS ambiguity, leading to confusion with VERBNOUN in particular. Example VERB$_3$ pairs incorrectly clustered are: (*study*, *studies*), (*run*, *runs*), and (*like*, *likes*). This polysemy results in the distance represented in the DIFFVEC for such pairs being above average for VERB$_3$, and consequently clustered with other cross-POS relations.

For VERB$_{\text{Past}}$, a single relatively pure cluster was generated, with minor contamination due to pairs such as (*hurt*, *saw*), (*utensil*, *saw*), and (*wipe*, *saw*). Here, the noun saw is ambiguous with a high-frequency past-tense verb; hurt and wipe also have ambigous POS.

A related phenomenon was observed for NOUN$_{\text{Coll}}$, where the instances were assigned to a large mixed cluster containing word pairs where the second word referred to an animal, reflecting the fact that most of the collective nouns in our dataset relate to animals, e.g. (*stand*, *horse*), (*ambush*, *tigers*), (*antibiotics*, *bacteria*). This is interesting from a DIFFVEC point of view, since it shows that the lexical semantics of one word in the pair can overwhelm the semantic content of the DIFFVEC (something that we return to investigate in Section 3.2.4.4). LEXSEM$_{\text{Mero}}$ was also split into multiple clusters along topical lines, with separate clusters for weapons, dwellings, vehicles, etc.

Given the encouraging results from our clustering experiment, we next evaluate DIFFVECs in a supervised relation classification setting.

3.2.4 Classification

A natural question is whether we can accurately characterise lexical relations through supervised learning over the DIFFVECs. For these experiments we use the w2v, w2v$_{\text{wiki}}$, FT$_{\text{cbow}}$, FT$_{\text{sg}}$ and SVD$_{\text{wiki}}$ embeddings exclusively (based on their superior performance in the clustering experiment), and a subset of the relations which is both representative of the breadth of the full relation set, and for which we have sufficient data for supervised training and evaluation, namely: NOUN$_{\text{Coll}}$, LEXSEM$_{\text{Event}}$, LEXSEM$_{\text{Hyper}}$, LEXSEM$_{\text{Mero}}$, NOUN$_{\text{SP}}$, PREFIX, VERB$_3$, VERB$_{\text{3Past}}$, and VERB$_{\text{Past}}$ (see Tables 3.2 and 3.3).

We consider two applications: (1) a CLOSED-WORLD setting similar to the unsupervised evaluation, in which the classifier only encounters word pairs which correspond to one of the nine relations; and (2) a more challenging OPEN-WORLD setting where random word pairs — which may or may not correspond to one of our relations — are included in the evaluation. For both settings, we further investigate whether there is a lexical memorisation effect for

a broad range of relation types of the sort identified by Weeds et al. (2014) and Levy et al. (2015b) for hypernyms, by experimenting with disjoint training and test vocabulary.

3.2.4.1 CLOSED-WORLD Classification

For the CLOSED-WORLD setting, we train and test a multiclass classifier on datasets comprising $\langle$DIFFVEC, $r\rangle$ pairs, where r is one of our nine relation types, and DIFFVEC is based on one of w2v, w2v_{wiki}, FT_{cbow}, FT_{sg} and SVD. As a baseline, we cluster the data as described in Section 3.2.3, running the clusterer several times over the 9-relation data to select the optimal V-Measure value based on the development data, resulting in 50 clusters. We label each cluster with the majority class based on the training instances, and evaluate the resultant labelling for the test instances.

We use an SVM with a linear kernel, and report results from 10-fold cross-validation in Table 3.5.

The SVM achieves a higher F-score than the baseline on almost every relation, particularly on $\text{LEXSEM}_{\text{Hyper}}$, and the lower-frequency NOUN_{SP}, $\text{NOUN}_{\text{Coll}}$, and PREFIX. Most of the relations — even the most difficult ones from our clustering experiment such as $\text{LEXSEM}_{\text{Hyper}}$, $\text{LEXSEM}_{\text{Event}}$, and $\text{LEXSEM}_{\text{Mero}}$— are classified with very high F-score. That is, with a simple linear transformation of the embedding dimensions, we are able to achieve near-perfect results. The PREFIX relation achieved markedly lower recall, resulting in a lower F-score, due to large differences in the predominant usages associated with the respective words (e.g., (*union*, *reunion*), where the vector for union is heavily biased by contexts associated with trade unions, but reunion is heavily biased by contexts relating to social get-togethers; and (*entry*, *reentry*), where entry is associated with competitions and entrance to schools, while reentry is associated with space travel). Somewhat surprisingly, given the small dimensionality of the input (vectors of size 300 for all three methods), we found that the linear SVM slightly outperformed a non-linear SVM using an RBF kernel. We observe no real difference between w2v_{wiki} and SVD_{wiki}, supporting the hypothesis of Levy et al. (2015a) that under appropriate parameter settings, count-based methods achieve high results. The impact of the training data volume for pre-training of the embeddings is also

Relation	**Baseline**	**w2v**	**w2v$_{wiki}$**	**SVD$_{wiki}$**	**FT$_{sg}$**	**FT$_{cbow}$**
LEXSEM$_{Hyper}$	0.60	0.93	0.91	0.91	0.93	0.90
LEXSEM$_{Mero}$	0.90	0.97	0.96	0.96	0.97	0.97
LEXSEM$_{Event}$	0.87	0.98	0.97	0.97	0.98	0.98
NOUN$_{SP}$	0.00	0.83	0.78	0.74	0.80	0.80
VERB$_{3}$	0.99	0.98	0.96	0.97	0.99	0.97
VERB$_{Past}$	0.78	0.98	0.98	0.95	0.98	0.97
VERB$_{3Past}$	0.99	0.98	0.98	0.96	1.00	1.00
PREFIX	0.00	0.82	0.34	0.60	0.80	0.78
NOUN$_{Coll}$	0.19	0.95	0.91	0.92	0.93	0.91

Table 3.5 F-scores ($\mathscr{F}$) for CLOSED-WORLD classification, for a baseline method based on clustering + majority-class labelling, a multiclass linear SVM trained on w2v, w2v$_{wiki}$, SVD$_{wiki}$, FT$_{sg}$ and FT$_{cbow}$ DIFFVEC inputs.

less pronounced than in the case of our clustering experiment. In addition, we also see that FT$_{cbow}$ and FT$_{sg}$ achieve results close to w2v even thought they are trained on less amount of data.

3.2.4.2 OPEN-WORLD Classification

We now turn to a more challenging evaluation setting: a test set including word pairs drawn at random. This setting aims to illustrate whether a DIFFVEC-based classifier is capable of differentiating related word pairs from noise, and can be applied to open data to learn new related word pairs.[10]

For these experiments, we train a binary classifier for each relation type, using $\frac{2}{3}$ of our relation data for training and $\frac{1}{3}$ for testing. The test data is augmented with an equal quantity of random pairs, generated as follows:

(1) sample a seed lexicon by drawing words proportional to their frequency in Wikipedia;[11]

(2) take the Cartesian product over pairs of words from the seed lexicon;

(3) sample word pairs uniformly from this set.

This procedure generates word pairs that are representative of the frequency profile of our corpus.

[10]Hereafter we provide results for w2v, FT$_{cbow}$, and FT$_{sg}$ only.

[11]Filtered to consist of words for which we have embeddings.

Relation	Orig			+neg		
	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{F}$	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{F}$
LEXSEM$_{\text{Hyper}}$	0.95	**0.92**	**0.93**	**0.99**	0.84	0.91
LEXSEM$_{\text{Mero}}$	0.13	**0.96**	0.24	**0.95**	0.84	**0.89**
LEXSEM$_{\text{Event}}$	0.44	**0.98**	0.61	**0.93**	0.90	**0.91**
NOUN$_{\text{SP}}$	0.95	**0.68**	0.8	**1.00**	**0.68**	**0.81**
VERB$_3$	0.75	**1.00**	0.86	**0.93**	0.93	**0.93**
VERB$_{\text{Past}}$	0.94	**0.86**	**0.90**	**0.97**	0.84	**0.90**
VERB$_{\text{3Past}}$	0.76	**0.95**	0.84	**0.87**	0.93	**0.90**
PREFIX	**1.00**	**0.29**	**0.44**	**1.00**	0.13	0.23
NOUN$_{\text{Coll}}$	0.43	**0.74**	0.55	**0.97**	0.41	**0.57**

Table 3.6 Precision ($\mathscr{P}$) and recall ($\mathscr{R}$) for OPEN-WORLD classification, using the binary classifier without ("Orig") and with ("+neg") negative samples.

We train 9 binary RBF-kernel SVM classifiers on the training partition, and evaluate on our randomly augmented test set.[12] Fully annotating our random word pairs is prohibitively expensive, so instead, we manually annotated only the word pairs which were positively classified by one of our models. The results of our experiments with `wordvec` and `fasttext` models are presented in the left half of Tables 3.6, 3.7, 3.8, in which we report on results over the combination of the original test data from Section 3.2.4.1 and the random word pairs, noting that recall ($\mathscr{R}$) for OPEN-WORLD takes the form of relative recall (Pantel et al., 2004) over the positively-classified word pairs. The results are much lower than for the closed-word setting (Table 3.5), most notably in terms of precision ($\mathscr{P}$). For instance, the random pairs (*have*, *works*), (*turn*, *took*), and (*works*, *started*) were incorrectly classified as VERB$_3$, VERB$_{\text{Past}}$ and VERB$_{\text{3Past}}$, respectively. That is, the model captures syntax, but lacks the ability to capture lexical paradigms, and tends to overgeneralise.

3.2.4.3 OPEN-WORLD Training with Negative Sampling

To address the problem of incorrectly classifying random word pairs as valid relations, we retrain the classifier on a dataset comprising both valid and automatically-generated negative distractor samples. The basic intuition behind this approach is to construct samples which

[12]The gamma parameter of the RBF-kernel was optimised by doing a grid-search.

Relation	Orig			+neg		
	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{F}$	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{F}$
LEXSEM$_{\text{Hyper}}$	**0.96**	0.65	0.78	**0.96**	**0.78**	**0.86**
LEXSEM$_{\text{Mero}}$	0.23	**0.95**	0.38	**0.94**	0.80	**0.86**
LEXSEM$_{\text{Event}}$	0.87	**0.96**	**0.91**	**0.93**	0.86	0.89
NOUN$_{\text{SP}}$	0.71	**0.60**	0.65	**0.79**	**0.60**	**0.68**
VERB$_{3}$	0.87	**0.87**	**0.87**	**0.97**	0.78	0.86
VERB$_{\text{Past}}$	0.89	**0.89**	0.89	**0.96**	**0.89**	**0.93**
VERB$_{\text{3Past}}$	0.75	**0.94**	**0.83**	**0.86**	0.78	0.81
PREFIX	**1.00**	**0.50**	0.67	**1.00**	0.43	0.60
NOUN$_{\text{Coll}}$	0.40	**0.83**	0.54	**0.82**	0.51	**0.63**

Table 3.7 Precision ($\mathscr{P}$) and recall ($\mathscr{R}$) for OPEN-WORLD classification, using the binary classifier without ("Orig") and with ("+neg") negative samples, FastText SG.

Relation	Orig			+neg		
	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{F}$	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{F}$
LEXSEM$_{\text{Hyper}}$	0.56	**0.74**	0.64	**0.93**	0.63	**0.75**
LEXSEM$_{\text{Mero}}$	0.19	**0.94**	0.33	**0.73**	0.77	**0.75**
LEXSEM$_{\text{Event}}$	0.50	**0.96**	0.66	**0.78**	0.80	**0.79**
NOUN$_{\text{SP}}$	0.31	**0.80**	0.45	**0.53**	0.64	**0.58**
VERB$_{3}$	0.50	**0.87**	0.63	**0.96**	0.76	**0.85**
VERB$_{\text{Past}}$	0.92	**0.86**	**0.89**	**0.96**	0.78	0.86
VERB$_{\text{3Past}}$	0.42	**0.91**	0.57	**0.72**	0.72	**0.72**
PREFIX	**0.62**	0.50	**0.55**	–	–	–
NOUN$_{\text{Coll}}$	0.54	**0.69**	**0.61**	**0.84**	0.47	0.60

Table 3.8 Precision ($\mathscr{P}$) and recall ($\mathscr{R}$) for OPEN-WORLD classification, using the binary classifier without ("Orig") and with ("+neg") negative samples, FastText CBoW.

will force the model to learn decision boundaries that more tightly capture the true scope of a given relation. To this end, we automatically generated two types of negative distractors:

opposite pairs: generated by switching the order of word pairs, $Oppos_{w1,w2}$ = $\mathbf{word_1} - \mathbf{word_2}$. This ensures the classifier adequately captures the asymmetry in the relations.

shuffled pairs: generated by replacing w_2 with a random word w_2' from the same relation, $Shuff_{w1,w2} = \mathbf{word_2'} - \mathbf{word_1}$. This is targeted at relations that take specific word classes in particular positions, e.g., (VB, VBD) word pairs, so that the model learns to encode the relation rather than simply learning the properties of the word classes.

Both types of distractors are added to the training set, such that there are equal numbers of valid relations, opposite pairs and shuffled pairs.

After training our classifier, we evaluate its predictions in the same way as in Section 3.2.4.2, using the same test set combining related and random word pairs.[13] The results are shown in the right half of Tables 3.6, 3.7, 3.8 (as "+neg"). Observe that the precision is much higher and recall somewhat lower compared to the classifier trained with only positive samples. This follows from the adversarial training scenario: using negative distractors results in a more conservative classifier, that correctly classifies the vast majority of the random word pairs as not corresponding to a given relation, resulting in higher precision at the expense of a small drop in recall. Overall this leads to higher F-scores, as shown in Figure 3.3, other than for hypernyms ($\text{LEXSEM}_{\text{Hyper}}$) and prefixes (PREFIX). For example, the standard classifier for $\text{NOUN}_{\text{Coll}}$ learned to match word pairs including an animal name (e.g., (*plague*, *rats*)), while training with negative samples resulted in much more conservative predictions and consequently much lower recall. The classifier was able to capture (*herd*, *horses*) but not (*run*, *salmon*), (*party*, *jays*) or (*singular*, *boar*) as instances of $\text{NOUN}_{\text{Coll}}$, possibly because of polysemy. The most striking difference in performance was for $\text{LEXSEM}_{\text{Mero}}$, where the standard classifier generated many false positive noun pairs (e.g. (*series*, *radio*)), but the false positive rate was considerably reduced with negative sampling.

3.2.4.4 Lexical Memorisation

Weeds et al. (2014) and Levy et al. (2015b) showed that supervised methods using DIFFVECs achieve artificially high results as a result of "lexical memorisation" over frequent words associated with the hypernym relation. For example, (*animal*, *cat*), (*animal*, *dog*), and (*animal*, *pig*) all share the superclass `animal`, and the model thus learns to classify as positive any word pair with `animal` as the first word.

[13] But noting that relative recall for the random word pairs is based on the pool of positive predictions from both models.

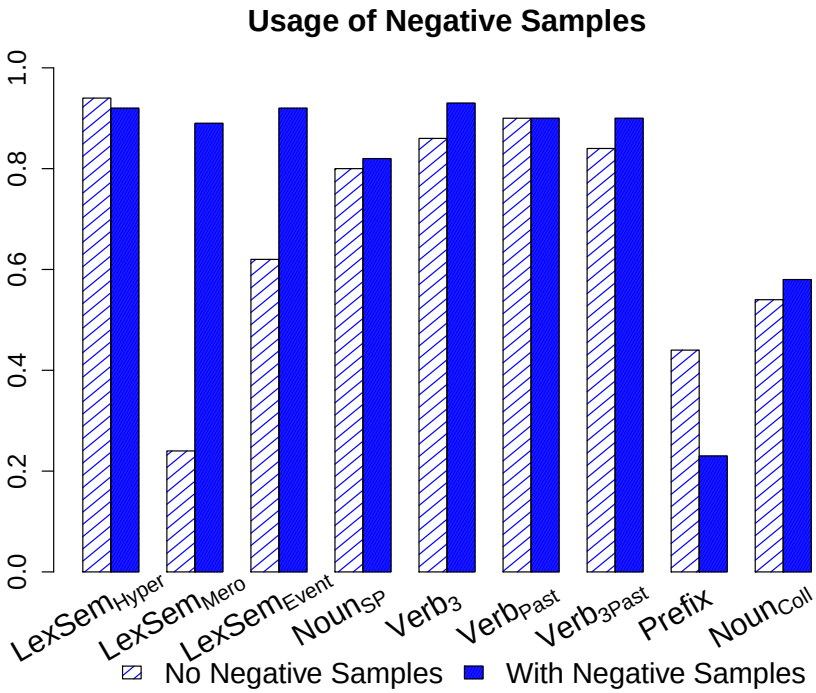


Figure 3.3 F-score for w2v OPEN-WORLD classification, comparing models trained with and without negative samples.

To address this effect, we follow Levy et al. (2015b) in splitting our vocabulary into training and test partitions, to ensure there is no overlap between training and test vocabulary. We then train classifiers with and without negative sampling (Section 3.2.4.3), incrementally adding the random word pairs from Section 3.2.4.2 to the test data (from no random word pairs to five times the original size of the test data) to investigate the interaction of negative sampling with greater diversity in the test set when there is a split vocabulary. The results are shown in Figure 3.4.

Observe that the precision for the standard classifier decreases rapidly as more random word pairs are added to the test data. In comparison, the precision when negative sampling is used shows only a small drop-off, indicating that negative sampling is effective at maintaining precision in an OPEN-WORLD setting even when the training and test vocabulary are disjoint. This benefit comes at the expense of recall, which is much lower when negative sampling is used (note that recall stays relatively constant as random word pairs are added, as the vast majority of them do not correspond to any relation). At the maximum level of random word

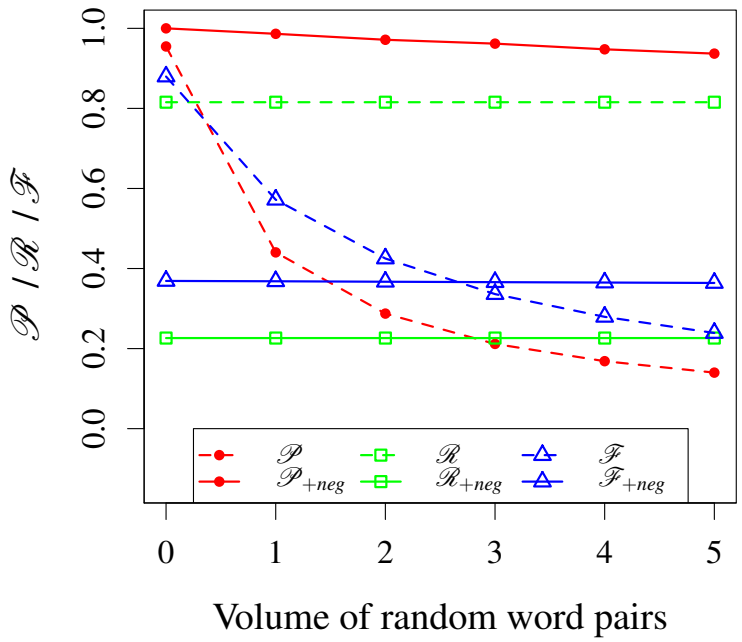


Figure 3.4 Evaluation of the OPEN-WORLD model when trained on split vocabulary, for varying numbers of random word pairs in the test dataset (expressed as a multiplier relative to the number of CLOSED-WORLD test instances).

pairs in the test data, the F-score for the negative sampling classifier is higher than for the standard classifier.

3.3 Language Modelling for Russian

We now turn to experiments with Russian, a more morphologically rich language. In order to evaluate the models, we first manually translate part of the initial dataset (the relations that were used in the classification task) and augment it with extra morphosyntactic relation types. The morphosyntactic pairs were automatically sampled from the UniMorph database. Tables 3.9 and 3.10 provide a more detailed information on the resulting dataset.

In there experiments, we only use `fasttext` embeddings since they demonstrated a superior performance in our experiments with English.

Relation	**Description**	**Example**
LEXSEM$_{\text{Hyper}}$	hypernym	(*jivotnoje*, *sobaka*)
LEXSEM$_{\text{Mero}}$	meronym	(*samolet*, *kokpit*)
LEXSEM$_{\text{Event}}$	object's action	(*zastegnut`*, *pal`to*)
NOUN$_{\text{SP}}$	NOUN;SG →NOUN;PL	(*god*, *gody*)
VERB$_{\text{1SgPrs}}$	INF → V;1SG;PRES	(*begat`*, *begaju*)
VERB$_{\text{1PlPrs}}$	INF → V;1PL;PRES	(*begat`*, *begajem*)
VERB$_{\text{2SgPrs}}$	INF → V;2SG;PRES	(*begat`*, *begaješ`*)
VERB$_{\text{2PlPrs}}$	INF → V;2PL;PRES	(*begat`*, *begajete*)
VERB$_{\text{3SgPrs}}$	INF → V;SG;PRES	(*begat`*, *begajet*)
VERB$_{\text{3PlPrs}}$	INF → V;3PL;PRES	(*begat`*, *begajut*)
VERB$_{\text{PlPast}}$	INF → V;PL;PAST	(*begat`*, *begali*)
VERB$_{\text{MascSgPast}}$	INF → V;M;SG;PAST	(*begat`*, *begal*)
VERB$_{\text{FemSgPast}}$	INF → V;F;SG;PAST	(*begat`*, *begala*)
VERB$_{\text{NeurSgPast}}$	INF → V;N;SG;PAST	(*begat`*, *begalo*)
PREFIX	prefixing with *re* morpheme	(*izbrat`*, *pereizbrat`*)
NOUN$_{\text{Coll}}$	collective noun	(*kolonija*, *murav`i*)

Table 3.9 Description of the 16 lexical relations for Russian language. Examples are translations of corresponding English instances (see continuation on Table 3.10).

3.3.1 Closed-World Classification

First, similar to the experiments with English, we test the models' ability to differentiate these relations in the CLOSED-WORLD classification setting.

The model's performance is quite high for morphosyntactic relations. Compared to English, Russian word forms are more indicative of morphosyntactic properties of the form and less ambiguous. Another possible cause might be a large overlap in the train and test lexicons, because many of the forms share the same lemma. In order to test the models' ability to generalise across multipe lemmata, we do an extra experiment where all word pairs in morphosyntatic relations are sampled from different lemmata (see Section 3.3.3). Generally, similar to English, we observe that in such a setting the model achieves good results on most types of relations. Higher accuracy on NOUN$_{\text{Coll}}$ is explained by the fact that Russian does not present such a fine-grained distinction of animal groups, assigning all of them to roughly five large categories.

Relation	Pairs	Source
LEXSEM$_{\text{Hyper}}$	1008	Translation of SemEval'12 + BLESS
LEXSEM$_{\text{Mero}}$	2520	Translation of SemEval'12 + BLESS
LEXSEM$_{\text{Event}}$	3282	Translation of BLESS
NOUN$_{\text{SP}}$	100	Translation of MSR
VERB$_{\text{1SgPrs}}$	100	UniMorph
VERB$_{\text{1PlPrs}}$	100	UniMorph
VERB$_{\text{2SgPrs}}$	100	UniMorph
VERB$_{\text{2PlPrs}}$	100	UniMorph
VERB$_{\text{3SgPrs}}$	100	UniMorph
VERB$_{\text{3PlPrs}}$	100	UniMorph
VERB$_{\text{PlPast}}$	100	UniMorph
VERB$_{\text{MascSgPast}}$	100	UniMorph
VERB$_{\text{FemSgPast}}$	100	UniMorph
VERB$_{\text{NeurSgPast}}$	100	UniMorph
PREFIX	113	Translation of English Wiktionary
NOUN$_{\text{Coll}}$	131	Russian Web source

Table 3.10 Number of samples and sources of the 16 lexical relations for Russian language.

Relation	$\texttt{FT}_{\text{sg}}$	$\texttt{FT}_{\text{cbow}}$
LEXSEM$_{\text{Hyper}}$	0.92	0.92
LEXSEM$_{\text{Mero}}$	0.96	0.96
LEXSEM$_{\text{Event}}$	0.99	0.99
NOUN$_{\text{SP}}$	0.69	0.85
VERB$_{\text{1SgPrs}}$	0.97	1.0
VERB$_{\text{1PlPrs}}$	1.0	1.0
VERB$_{\text{2SgPrs}}$	1.0	1.0
VERB$_{\text{2PlPrs}}$	1.0	1.0
VERB$_{\text{3SgPrs}}$	0.96	1.0
VERB$_{\text{3PlPrs}}$	0.98	1.0
VERB$_{\text{PlPast}}$	1.0	1.0
VERB$_{\text{MascSgPast}}$	1.0	1.0
VERB$_{\text{FemSgPast}}$	1.0	1.0
VERB$_{\text{NeurSgPast}}$	1.0	1.0
PREFIX	0.72	0.83
NOUN$_{\text{Coll}}$	0.97	0.98

Table 3.11 F-scores ($\mathscr{F}$) for CLOSED-WORLD classification a multiclass linear SVM trained on $\texttt{FT}_{\text{sg}}$ and $\texttt{FT}_{\text{cbow}}$ DIFFVEC inputs for Russian language.

3.3.2 Open-World Classification

Now we check the models' performance in a more realistic setting. Following our experiments with English data, we add noisy samples (i.e. pairs of unrelated words) to the dataset. As shown on Tables 3.12 and 3.13, the performance drops, especially in the case of lexical semantic relations. Similarly to English, LEXSEM$_{\text{Mero}}$ are affected the most. We also observe that FT$_{\text{sg}}$ in most cases outperforms FT$_{\text{cbow}}$, especially in the case of PREFIX.

Relation	**Orig**			**+neg**		
	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{F}$	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{F}$
LEXSEM$_{\text{Hyper}}$	0.45	**0.88**	0.59	**0.89**	0.78	**0.83**
LEXSEM$_{\text{Mero}}$	0.12	**0.97**	0.22	**0.92**	0.86	**0.89**
LEXSEM$_{\text{Event}}$	0.63	**0.99**	0.77	**0.91**	0.96	**0.94**
NOUN$_{\text{SP}}$	**0.96**	**0.71**	**0.82**	**0.96**	0.66	0.78
VERB$_{\text{1SgPrs}}$	1.0	0.86	0.92	1.0	0.86	0.92
VERB$_{\text{1PlPrs}}$	1.0	1.0	1.0	1.0	1.0	1.0
VERB$_{\text{2SgPrs}}$	1.0	1.0	1.0	1.0	1.0	1.0
VERB$_{\text{2PlPrs}}$	0.70	**1.0**	0.83	**0.95**	0.95	**0.95**
VERB$_{\text{3SgPrs}}$	0.96	0.92	0.94	**1.0**	**0.96**	**0.98**
VERB$_{\text{3PlPrs}}$	1.0	0.93	0.96	1.0	0.93	0.96
VERB$_{\text{PlPast}}$	1.0	**1.0**	**1.0**	1.0	0.96	0.98
VERB$_{\text{MascSgPast}}$	1.0	0.96	0.98	1.0	0.96	0.98
VERB$_{\text{FemSgPast}}$	0.95	**0.90**	**0.93**	**1.0**	0.86	0.92
VERB$_{\text{NeurSgPast}}$	1.0	**1.0**	**1.0**	1.0	0.82	0.90
PREFIX	1.0	**0.68**	**0.81**	1.0	0.47	0.64
NOUN$_{\text{Coll}}$	1.0	**0.88**	**0.94**	1.0	0.74	0.85

Table 3.12 Precision ($\mathscr{P}$) and recall ($\mathscr{R}$) for OPEN-WORLD FastText SG classification in Russian, using the binary classifier without ("Orig") and with ("+neg") negative samples.

As presented on Tables 3.12 and 3.13, negative sampling improves the results for lexical relations such as hypernymy, meronymy, and events, although morphosyntactic ones generally do not need it. This support are findings for English. For PREFIX and NOUN$_{\text{Coll}}$ it actually has a negative affect. In the case of collective nouns, we can address it to the fact that word shuffling within a single relation leads to many positive rather than negative examples (due to the specificity of Russian, mentioned earlier).

Relation	Orig			+neg		
	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{F}$	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{F}$
LEXSEM$_{\text{Hyper}}$	0.19	**0.86**	0.31	**0.84**	0.69	**0.76**
LEXSEM$_{\text{Mero}}$	0.16	**0.97**	0.27	**0.77**	0.76	**0.77**
LEXSEM$_{\text{Event}}$	0.55	**0.99**	0.71	**0.88**	0.92	**0.90**
NOUN$_{\text{SP}}$	0.32	**0.88**	0.47	**1.00**	0.67	**0.80**
VERB$_{\text{1SgPrs}}$	1.00	**0.90**	**0.95**	1.00	0.80	0.89
VERB$_{\text{1PlPrs}}$	1.00	**0.93**	**0.97**	1.00	0.83	0.91
VERB$_{\text{2SgPrs}}$	1.00	0.94	0.97	1.00	0.94	0.97
VERB$_{\text{2PlPrs}}$	**1.00**	0.58	0.73	0.60	**0.95**	0.73
VERB$_{\text{3SgPrs}}$	0.5	0.82	0.62	**0.91**	**0.88**	**0.89**
VERB$_{\text{3PlPrs}}$	1.00	**1.00**	**1.00**	1.00	0.96	0.98
VERB$_{\text{PlPast}}$	0.93	1.00	0.96	**1.00**	1.00	**1.00**
VERB$_{\text{MascSgPast}}$	0.97	0.91	0.93	**1.00**	0.91	**0.95**
VERB$_{\text{FemSgPast}}$	0.39	**0.97**	0.56	**1.00**	0.93	**0.97**
VERB$_{\text{NeurSgPast}}$	1.00	0.82	0.90	1.00	0.82	0.90
PREFIX	1.00	**0.21**	**0.35**	1.00	0.08	0.15
NOUN$_{\text{Coll}}$	1.00	**0.90**	**0.95**	1.00	0.73	0.84

Table 3.13 Precision ($\mathscr{P}$) and recall ($\mathscr{R}$) for OPEN-WORLD FastText CBoW classification in Russian, using the binary classifier without ("Orig") and with ("+neg") negative samples.

3.3.3 Split Volcabulary for Morphology

We additionally measure the models' ability to generalise across unseen lemmata for morphosyntactic relations and run extra experiments where the training and test lexicon for morphosyntactic relations do not overlap.[14] We report the performance in OPEN-WORLD settings and only for this type of relation (although we train on all training data). Tables 3.14 and 3.15 do not show a substantial drop of performance.

To summarise, we observe that morphosyntatic relations are captured better compared to other relaton types, and character-level models learn them more efficiently and need less amount of data. Accuracy for Russian is higher than for English which we addressed to more transparent form–meaning relation there. In the next section, we will further compare various character-level architectures in their ability to present morphological awareness in a machine translation task.

[14]The lexicons are the same for both models.

Relation	Orig			+neg		
	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{F}$	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{F}$
NOUN_{SP}	1.00	0.74	**0.85**	1.00	0.74	0.85
$\text{VERB}_{\text{1SgPrs}}$	1.0	**0.67**	0.81	1.0	0.64	0.78
$\text{VERB}_{\text{1PlPrs}}$	1.0	0.91	0.96	1.0	**0.92**	0.96
$\text{VERB}_{\text{2SgPrs}}$	1.0	**0.97**	**0.99**	1.0	0.92	0.95
$\text{VERB}_{\text{2PlPrs}}$	1.0	0.88	0.94	1.00	0.88	0.94
$\text{VERB}_{\text{3SgPrs}}$	1.0	0.88	0.94	**1.0**	0.88	0.94
$\text{VERB}_{\text{3PlPrs}}$	1.0	0.86	0.93	1.0	0.86	0.93
$\text{VERB}_{\text{PlPast}}$	1.0	**0.92**	**0.96**	1.0	0.87	0.93
$\text{VERB}_{\text{MascSgPast}}$	1.0	0.96	0.98	1.0	0.96	0.98
$\text{VERB}_{\text{FemSgPast}}$	0.95	0.78	**0.86**	0.95	0.75	0.84
$\text{VERB}_{\text{NeurSgPast}}$	1.0	**0.97**	**0.98**	1.0	0.94	0.96

Table 3.14 Precision ($\mathscr{P}$) and recall ($\mathscr{R}$) for OPEN-WORLD FastText SG classification in Russian with split lexicon, using the binary classifier without ("Orig") and with ("+neg") negative samples.

Relation	Orig			+neg		
	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{F}$	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{F}$
NOUN_{SP}	0.41	0.71	0.52	**0.88**	**0.74**	**0.80**
$\text{VERB}_{\text{1SgPrs}}$	1.00	0.72	0.84	1.00	0.72	0.84
$\text{VERB}_{\text{1PlPrs}}$	1.00	**0.93**	**0.97**	1.00	0.87	0.93
$\text{VERB}_{\text{2SgPrs}}$	1.00	0.96	0.98	1.00	0.96	0.98
$\text{VERB}_{\text{2PlPrs}}$	**0.96**	0.73	0.83	0.84	**0.94**	**0.89**
$\text{VERB}_{\text{3SgPrs}}$	0.52	**0.96**	0.67	**0.82**	0.93	**0.87**
$\text{VERB}_{\text{3PlPrs}}$	1.00	**0.95**	**0.97**	1.00	0.92	0.96
$\text{VERB}_{\text{PlPast}}$	0.91	**1.00**	0.95	**1.00**	0.96	**0.98**
$\text{VERB}_{\text{MascSgPast}}$	**1.00**	0.73	0.84	0.88	**0.84**	**0.86**
$\text{VERB}_{\text{FemSgPast}}$	0.81	**0.93**	0.87	**0.90**	0.87	**0.89**
$\text{VERB}_{\text{NeurSgPast}}$	1.00	**0.93**	**0.97**	1.00	0.87	0.93

Table 3.15 Precision ($\mathscr{P}$) and recall ($\mathscr{R}$) for OPEN-WORLD FastText CBoW classification in Russian, using the binary classifier without ("Orig") and with ("+neg") negative samples.

3.4 Machine Translation

Models of end-to-end machine translation based on neural networks can produce excellent translations, rivalling or surpassing traditional statistical machine translation systems (Bahdanau et al., 2015; Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014). A central challenge in neural MT is handling rare and uncommon words. Conventional neural MT

models use a fixed modest-size vocabulary, such that the identity of rare words is lost, which makes their translation exceedingly difficult. Accordingly, sentences containing rare words tend to be translated much more poorly than those containing only common words (Bahdanau et al., 2015; Sutskever et al., 2014). The rare word problem is exacerbated when translating from morphologically rich languages, where the large number of morphological variants of words result in a huge vocabulary with a heavy tail. For example in Russian, there are at least 70 word forms for dog, encoding case, gender, age, number, sentiment and other semantic connotations. Many of them share a common lemma, and contain regular morphological affixation; consequently much of the information required for translation is present, but not in an accessible form for models of neural MT.

In many cases, the OOV problem is addressed by incorporating character-level word representations largely belonging to one of two classes, namely convolutional neural networks (CNNs) and recurrent neural networks based on long-short term memory (LSTM) units (Hochreiter and Schmidhuber, 1997). But there has been no investigation of what each of the models captures and how well they can model morphology in particular. In this section, we fill this gap by evaluating encoder-level representations of OOV words. To get the representations, we incorporate LSTM and CNN word representation models into two types of attentional machine translation models. Our evaluation includes both intrinsic and extrinsic metrics, where we compare these approaches based on their translation performance as well as their ability to recover synonyms for the rare words. Intrinsic analysis shows that there are only minor differences in translation performance, although detailed analysis shows that the character-based LSTM is overal best at capturing morphological regularities.

3.4.1 Models

Now we turn to the problem of learning word representations. We consider character-level encoding methods which we compare to the baseline word embedding approach. We test three types of character representations: recurrent neural networks (RNNs) with LSTM units, convolutional neural networks (CNNs), and initialising source-level word embedding with pre-trained FastText embeddings.

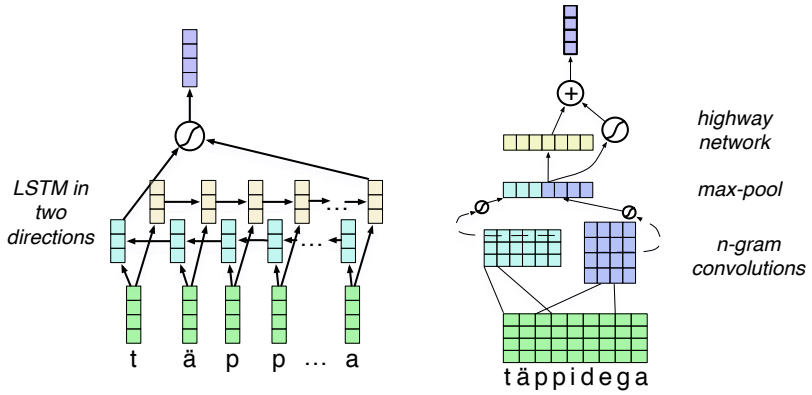


Figure 3.5 Model architecture for the several approaches to learning word representations, showing from left: BiLSTM over characters and the character convolution.

For CNN and RNN character encoders, we learn two word representations: one estimated from the characters, and a word embedding.[15] Then we run max pooling over both embeddings to obtain the word representation, $\mathbf{r}_w = \mathbf{m}_w \oslash \mathbf{e}_w$, where $\mathbf{m}_w$ is the embedding of word w and $\mathbf{e}_w$ is the sub-word encoding. The max pooling operation $\oslash$ captures non-compositionality in the semantic meaning of a word relative to its sub-parts. We hypothesise that the model will favour unit-based embeddings for rare words and word-based for more common ones.

Each word is expressed with its constituent units as follows. Let $\mathscr{U}$ be the vocabulary of sub-word units, i.e., characters, E_u be the dimensionality of unit embeddings, and $M \in \mathbb{R}^{E_u \times |\mathscr{U}|}$ be the matrix of unit embeddings. Suppose that a word w from the source dictionary is made up of a sequence of units $\mathscr{U}_w := [u_1, \ldots, u_{|w|}]$, where $|w|$ stands for the number of constituent units in the word. The resulting word representations are then fed to both attentional models as the source word embeddings.

Bidirectional LSTM Encoder The encoding of the word is formulated using a pair of LSTMs (denoted *biLSTM*) one operating left-to-right over the input sequence and another operating right-to-left, $\mathbf{h}_j^{\rightarrow} = \text{LSTM}(\mathbf{h}_{j-1}^{\rightarrow}, \mathbf{m}_{u_j})$ and $\mathbf{h}_j^{\leftarrow} = \text{LSTM}(\mathbf{h}_{j+1}^{\leftarrow}, \mathbf{m}_{u_j})$ where $\mathbf{h}_j^{\rightarrow}$ and

[15] We only include word embeddings for common words; rare words share an `UNK` embedding.

$\mathbf{h}_j^{\leftarrow}$ are the LSTM hidden states.[16] These are fed into a perceptron with a single hidden layer and a tanh activation function to form the word representation, $\mathbf{e}_w = \text{MLP}\left(\mathbf{h}_{|\mathscr{U}_w|}^{\rightarrow}, \mathbf{h}_1^{\leftarrow}\right)$.

Convolutional Encoder Another word encoder we consider is a convolutional neural network, inspired by a similar approach in language modelling (Kim et al., 2016). Let $U_w \in \mathbb{R}^{E_u \times |\mathscr{U}|_w}$ denote the unit-level representation of w, where the j-th column corresponds to the unit embedding of u_j. The idea of unit-level CNN is to apply a *kernel* $Q_l \in \mathbb{R}^{E_u \times k_l}$ with width k_l to U_w to obtain a feature map $\mathbf{f_l} \in \mathbb{R}^{|\mathscr{U}_w| - k_l + 1}$. More formally, for the j-th element of the feature map the convolutional representation is

$$\mathbf{f_l}(j) = \tanh(\langle U_{w,j}, Q_l \rangle + b)$$

where $U_{w,j} \in \mathbb{R}^{E_u \times k_l}$ is a slice from U_w which spans the representations of the j-th unit and its preceding $k_l - 1$ units, and

$$\langle A, B \rangle = \sum_{i,j} A_{ij} B_{ij} = \text{Tr}\left(AB^T\right)$$

denotes the Frobenius inner product. For example, suppose that the input has size $[4 \times 9]$, and a kernel has size $[4 \times 3]$ with a sliding step being 1. Then, we obtain a $[1 \times 7]$ feature map. This process implements a character n-gram, where n is equal to the width of the filter. The word representation is then derived by max pooling the feature maps of the kernels:

$$\forall l: \quad \mathbf{r}_w(l) = \max_j \mathbf{f_l}(j)$$

In order to capture interactions between the character n-grams obtained by the filters, a *highway network* (Srivastava et al., 2015) is applied after the max pooling layer,

$$\mathbf{e}_w = \mathbf{t} \odot \text{MLP}(\mathbf{r}_w) + (1 - \mathbf{t}) \odot \mathbf{r}_w,$$

[16] The memory cells are computed as part of the recurrence, suppressed here for clarity.

Set		**Ru-En**	**Et-En**
Train	Tokens	1,639K-1,809K	1,411K-1,857K
	Types	145K-65K	90K-25K
Development	Tokens	150K-168K	141K-188K
	Types	35K-18K	21K-9K
Test	Tokens	150K-167K	142K-189K
	Types	35K-18K	21K-8K
OOV	Types	45%	45%

Table 3.16 Corpus statistics for parallel data between Russian/Estonian and English. The OOV rate are the fraction of word types in the source language that are in the test set but are below the frequency cut-off or unseen in training.

where $\mathbf{t} = \text{MLP}_\sigma(\mathbf{r}_w)$ is a sigmoid gating function which modulates between a tanh MLP transformation of the input (left component) and preserving the input as is (right component).

In addition, we also run an experiment with **pre-trained `fasttext` embeddings** $\texttt{FT}_{\text{cbow}}$. We initialise source language embeddings and keep them constant during the training phrase, leaving everything else unchanged.

3.4.2 Experiments

Datasets. We use parallel bilingual data from Europarl for Estonian-English (Koehn, 2005), and web-crawled parallel data for Russian-English (Antonova and Misyurev, 2011). For preprocessing, we tokenise, lower-case, and filter out sentences longer than 30 words. We apply a frequency threshold of 5, replacing low-frequency words with a special UNK token. Table 3.16 presents the corpus statistics.

Extrinsic Evaluation: MT We apply the character-level models in the encoder of the neural attentional (Bahdanau et al., 2015) ("AM", soft-attentional) and neural operation sequence (Vylomova et al., 2016, v.1) ("OSM", hard-attentional) models, replacing the source word embedding component with either `BiLSTM` or `CNN` over characters, or `fasttext` pre-trained embeddings. To evaluate translations, we re-ranked the 100-best output translations from Moses[17] using the attentional models. The re-ranker includes standard features from

[17] `https://github.com/moses-smt`.

Language \ Model	Ru-En	Et-En
Phrase-based baseline	15.02	24.40
CHAR BILSTM$_{am}$	16.01	26.34
FASTTEXT$_{am}$	15.94	26.34
CHAR BILSTM$_{osm}$	15.81	26.14
CHAR CNN$_{am}$	15.90	26.14
CHAR CNN$_{osm}$	15.94	25.97
WORD$_{am}$	15.93	26.33
WORD$_{osm}$	15.70	26.03

Table 3.17 BLEU scores for re-ranking the test sets.

Moses plus extra feature(s) for each of the models. For AM we supply the log probability of the candidate translation, and for OSM we add two extra features corresponding to the generated alignment and the translation probabilities. The weights of the re-ranker are then trained using MERT (Och, 2003) with 100 restarts to optimise BLEU.

Table 3.17 presents BLEU score results. As seen, re-ranking based on the neural model scores outperforms the phrase-based baseline. However, the translation quality of the neural models are not significantly different. We assume that this is due to re-ranking of Moses translations rather than decoding. Also note that here we do not address the problem of OOV on the decoding side. Also, we observe that usage of pre-trained source embeddings does not lead to a significant improvement in such a setting.

Intrinsic Evaluation We now take a closer look at the embeddings learned by the models, based on how well they capture the *semantic* and *morphological* information in the nearest neighbour words. Learning representations for low frequency words is harder than that for high-frequency words, since low frequency words cannot capitalise as reliably on their contexts. Therefore, we split the test lexicon into 6 parts according to their frequency in the training set. Since we set the word frequency threshold to 5 for the training set, all words appearing in the lowest frequency band [0,4] are OOVs for the test set. For each word of the test set, we take its top-20 nearest neighbours from the whole training lexicon using cosine similarity.

Semantic Evaluation. We investigate how well the nearest neighbours are interchangeable with a query word in the translation process. So we formalise the notion of semantics of the source words based on their translations in the target language. We use *pivoting* to define the probability of a candidate word e' to be the synonym of the query word e, $p(e'|e) = \sum_f p(f|e)p(e'|f)$, where f is a target language word, and the translation probabilities inside the summation are estimated using a word-based translation model trained on the entire initial bilingual corpora. We then take the top-5 most probable words as the gold synonyms for each query word of the test set.[18]

We measure the quality of predicted nearest neighbours using multi-label accuracy[19] $\frac{1}{|S|}\sum_{w\in S}\mathbf{1}_{[\mathrm{G}(w)\cap\mathrm{N}(w)\neq\varnothing]}$ where $G(w)$ and $N(w)$ are the sets of gold standard synonyms and nearest neighbors for w respectively; the function $\mathbf{1}_{[C]}$ evaluates to one if the condition C is true, and zero otherwise. In other words, it is the fraction of words in S whose nearest neighbours and gold standard synonyms have non-empty overlap.

Table 3.18 presents the semantic evaluation results. As seen, for the *vanilla* (soft) attentional, model word- and character-level representations perform quite similar. In the case of the *hard* attentional model, CHAR $\mathrm{CNN}_{\mathrm{osm}}$ outperforms other representations by a large margin.

Morphological Evaluation. We now turn to evaluating the morphological component. We only focus on Russian since it has notoriously hard morphology. We run another morphological analyser, *mystem* (Segalovich, 2003), to generate *linguistically tagged* morphological analyses for a word, e.g. POS tags, case, person, plurality, etc. We represent each morphological analysis with a bit vector, where each 1 bit indicates the presence of a specific grammatical feature. Each word is then assigned a set of bit vectors corresponding to the set of its morphological analyses. As the *morphology similarity* between two words, we take the maximum of Hamming similarity[20] between the corresponding two sets of bit vectors.

[18] We remove query words whose frequency is less than a threshold in the initial bilingual corpora, since pivoting may not result in high quality synonyms for such words.

[19] We evaluated using mean reciprocal rank (MRR) measure as well, and obtained results consistent with the multi-label accuracy

[20] The Hamming similarity is the number of bits having the same value in two given bit vectors.

Model \ Freq.	0-4	5-9	10-14	15-19	20-50	50+
Russian						
WORD$_{am}$	–	0.32	0.52	0.65	**0.81**	**0.95**
WORD$_{osm}$	–	0.36	0.49	0.61	0.76	0.91
CHAR BILSTM$_{am}$	0.21	0.33	0.49	0.58	0.71	0.85
CHAR BILSTM$_{osm}$	0.16	0.34	0.48	0.59	0.71	0.85
CHAR CNN$_{am}$	0.13	0.23	0.38	0.47	0.61	0.84
CHAR CNN$_{osm}$	**0.43**	**0.71**	**0.77**	**0.77**	**0.81**	0.81
Estonian						
WORD$_{am}$	–	0.39	0.53	0.63	0.72	0.88
WORD$_{osm}$	–	0.48	0.62	0.70	**0.79**	**0.90**
CHAR BILSTM$_{am}$	0.12	0.30	0.37	0.45	0.52	0.70
CHAR BILSTM$_{osm}$	0.13	0.39	0.48	0.55	0.63	0.78
CHAR CNN$_{am}$	0.12	0.25	0.33	0.42	0.52	0.75
CHAR CNN$_{osm}$	**0.48**	**0.70**	**0.75**	**0.76**	0.78	0.78

Table 3.18 Semantic evaluation of nearest neighbours using multi-label accuracy on words in different frequency bands.

Table 3.19(a) shows the average morphology similarity between the words and their nearest neighbours across the frequency bands. Likewise, we represent the words based on their lemma features; Table 3.19(b) shows the average lemma similarity.

Table 3.20 lists the top five nearest neighbours for OOV words produced by the OSM models. BiLSTMs better capture morphological similarities expressed in suffixes and prefixes. We assume this is due to the fact that they are naturally biased towards the most recent inputs. CNNs, on the other hand, are more invariant of character positions and provide whole-word similarity.

3.5 Conclusion

We first evaluated embeddings obtained from language modelling and tested the generalisability of the vector difference approach across a broad range of lexical relations (in raw number and also variety) in English and Russian. Using clustering we showed that many types of morphosyntactic and morphosemantic differences are captured by DIFFVECs, but that lexical semantic relations are captured less well, a finding which is consistent with

Model \ Freq.	**0-4**	**5-9**	**10-14**	**15-19**	**20-50**	**50+**
WORD_{am}	-	0.70	0.73	075	0.78	0.82
WORD_{osm}	-	0.74	0.77	0.78	0.81	0.84
$\text{CHAR BILSTM}_{\text{am}}$	0.90	0.82	0.83	0.83	0.84	0.82
$\text{CHAR BILSTM}_{\text{osm}}$	**0.91**	**0.84**	**0.85**	**0.85**	**0.86**	**0.86**
$\text{CHAR CNN}_{\text{am}}$	0.82	0.76	0.77	0.78	0.79	0.81
$\text{CHAR CNN}_{\text{osm}}$	0.79	0.80	0.79	0.79	0.79	0.79

(a)

Model \ Freq.	**0-4**	**5-9**	**10-14**	**15-19**	**20-50**	**50+**
WORD_{am}	-	0.02	0.04	0.07	0.11	0.18
WORD_{osm}	-	0.03	0.05	0.06	0.09	0.15
$\text{CHAR BILSTM}_{\text{am}}$	0.08	0.06	0.10	0.11	0.12	0.21
$\text{CHAR BILSTM}_{\text{osm}}$	0.05	0.05	0.08	0.10	0.13	0.18
$\text{CHAR CNN}_{\text{am}}$	0.04	0.02	0.05	0.06	0.1	0.15
$\text{CHAR CNN}_{\text{osm}}$	**0.20**	**0.37**	**0.41**	**0.42**	**0.44**	**0.41**

(b)

Table 3.19 Morphology analysis for nearest neighbours based on (a) Grammar tag features, and (b) Lemma features, evaluated on Russian.

previous work (Köper et al., 2015). In contrast, classification over the DIFFVECs works extremely well in a closed-world setting, showing that dimensions of DIFFVECs encode lexical relations. Classification performs less well over open data, although with the introduction of automatically-generated negative samples, the results improve substantially. Negative sampling also improves classification when the training and test vocabulary are split to minimise lexical memorisation. Our comparison of word- and character-level models showed that the latter are able to achieve higher accuracy with less data using it more efficiently. We contrasted two character-level models, $\texttt{FT}_{\text{sg}}$ and $\texttt{FT}_{\text{cbow}}$, showing a superior performance of $\texttt{FT}_{\text{sg}}$ on this task in both English and Russian. Overall, we conclude that the DIFFVEC approach has impressive utility over a broad range of lexical relations, especially under supervised classification and morphosyntatic relations, presenting more regularity in general, are captured better than morphosemantic and lexical semantic, especially when we apply character-level models to morphologically rich languages.

In the MT task, we studied two types of attentional models augmented by `CNN`, `LSTM` and `fasttext` word embeddings. Our experiments on translation from Russian and Estonian

Ras+po+*lag*+a+jušč+ej *Disposing (inpraes,dat,sg,partcp,plen,f,ipf,intr)*	
CHAR CNN_{osm}	CHAR $BILSTM_{osm}$
ras+po+lag+a+jušč+iy *disposing (inpraes,nom,sg,partcp,plen,m,ipf,inan,intr)*	ras+slab+l+ja+ušč+ej *relaxing (inpraes,dat,sg,partcp,plen,f,ipf)*
ras+po+lag+a+jušč+im *disposing (inpraes,ins,sg,partcp,plen,m,ipf,intrn)*	so+pro+voj+d+a+jušč+ej *accompanying (inpraes,dat,sg,partcp,plen,f,ipf,tran)*
ras+po+lag+a+jušč+ije *disposing (inpraes,nom,pl,partcp,plen,ipf,intr)*	ras+slab+l+ja+ušč+uju *relaxing (inpraes,acc,sg,partcp,plen,f,ipf)*
ras+po+lag+a+jušč+ix *disposing (inpraes,gen,pl,partcp,plen,ipf,intr)*	ras+po+lag+a+jušč+iy *disposing (inpraes,nom,sg,partcp,plen,m,ipf,inan,intr)*
ras+po+lag+a+jušč+i+e+sja *disposing (inpraes,nom,pl,partcp,plen,ipf,act)*	pro+dvig+a+jušč+ej *promoting (inpraes,dat,sg,partcp,plen,f,ipf,act)*
S+*konfigur*+ir+ova+t` *Configure (v,pf,tran,inf)*	
CHAR CNN_{osm}	CHAR $BILSTM_{osm}$
s+konfigur+ir+ui+te *configure (v,pf,tran,pl,imper,2p)*	konfigur+ir+ova+t` *configure (v,ipf,tran,inf)*
s+konfigur+ova+li *configured (v,pf,tran,praet,pl,indic)*	s+korrekt+ir+ova+t` *adjust (v,pf,tran,inf)*
s+konfigur+ova+n *configured (v,pf,tran,praet,sg,partcp,brev,m,pass)*	s+*koordin*+ir+ova+t` *coordinate (v,pf,tran,inf)*
s+konstru+ir+ova+t` *construct (v,pf,tran,inf)*	s+fokus+ir+ova+t` *focus (v,pf,tran,in)*
s+kompil+ir+ova+t` *compile (v,pf,tran,inf)*	s+kompil+ir+ova+t` *compile (v,pf,tran,inf)*

Table 3.20 Analysis of the five most similar Russian words (initial word is OOV), under the CHAR CNN_{osm} and CHAR $BILSTM_{osm}$ word encodings based on cosine similarity. The diacritic ´ indicates softness. **POS tags**: *s*-noun, *a*-adjective, *v*-verb; **Gender**: *m*-masculine, *f*-feminine, *n*-neuter; **Number**: *sg*-singular, *pl*-plural; **Case**: *nom*-nominative, *gen*-genitive, *dat*-dative, *acc*-accusative, *ins*-instrumental, *abl*-prepositional, *loc*-locative; **Tense**: *praes*-present, *inpraes*-continuous, *praet*-past, *pf*-perfect, *ipf*-imperfect; *indic*-indicative; **Transitivity**: *trans*-transitive, *intr*-intransitive; **Adjective form**: *br*-brevity, *plen*-full form, *poss*-possessive; **Comparative**: *supr*-superlative, *comp*-comparative; **Noun person**: *1p*-first, *2p*-second, *3p*-third;

into English demonstrated that representation of out-of-vocabulary words with their sub-word units on the source side did not lead to a significant improvement in overall quality of machine translation. However, LSTMs applied to character sequences are more capable of learning morphological patterns in Russian and Estonian. Moreover, a hard attention mechanism leads to better capturing semantic and morphological regularities.

Chapter 4

Inflectional Morphology Models

4.1 Introduction

In this section, we focus on evaluation of neural models in terms of their grammar and syntax awareness. First, we aim to understand how much information about a word's morphosyntatic properties can be inferred directly from its sentential context and to what extent we are able to predict them. Second, we also evaluate morphosyntactic categories themselves in terms of their contextual predictability.

The primary evaluation for most contemporary language and translation modelling research is perplexity (Jelinek et al., 1977), BLEU (Papineni et al., 2002) or METEOR (Banerjee and Lavie, 2005). Undoubtedly, such metrics are necessary for extrinsic evaluation and comparison. However, relatively few studies have focused on intrinsic evaluation of grammaticality. Recently, Linzen et al. (2016) investigated the ability of an LSTM language model to capture sentential structure, by evaluating subject–verb agreement with respect to number. They show that under strong supervision, the LSTM is able to approximate structure-sensitive dependencies and achieves a surprisingly low error rate. In the same spirit, Belinkov et al. (2017) conduct an extensive analysis of neural machine translation models. The authors evaluate word representations at each layer, comparing those in both the encoder and decoder. Although the authors do not directly assess grammaticality, they

provide significant insights on what type of information is captured at each layer and how the target language might affect the source-side representation quality.

NLP systems are often required to generate grammatical text—for example, in machine translation, summarisation, dialogue, or grammar correction. Much work has been done on grammatical error correction: the study of systems that help non-native speakers correct their errors and provide them with feedback (Chodorow and Leacock, 2000; Dale and Kilgarriff, 2011; Leacock et al., 2014). For example, Sakaguchi et al. (2017) propose a dependency parsing scheme to repair ungrammatical sentences. Much work has also focused on the task of joint dependency parsing and disfluency detection (Honnibal and Johnson, 2014; Rasooli and Tetreault, 2014; Wu et al., 2015; Yoshikawa et al., 2016). Dahlmeier and Ng (2012) proposed a correction decoder that assigns a probability to a token as being ungrammatical, and then applied an error-type specific classifier to the ones with the highest values to make corrections. Ng et al. (2014) extended the task to whole-sentence correction, i.e. correcting all types of errors within a sentence. Another strand of work is solely devoted to parsing of ungrammatical sentences by adapting a parser to a specific domain (Berzak et al., 2016; Cahill, 2015; Nagata and Sakaguchi, 2016; Petrov and McDonald, 2012).

One component of grammaticality is the deployment of contextually appropriate closed-class morphemes. Here we introduce a novel self-contained task, **contextual inflection**, where a system must morphologically tag and inflect lemma tokens in sentential context. For example, in English, the system must reconstruct the correct word sequence *Moira proudly teaches two subjects* from the lemma sequence `Moira proudly teach two subject`. Among other things, this requires: (1) identifying `teach` as a verb in this context, (2) recognizing that `teach` should be inflected as 3rd-person singular to agree with the nearby noun, and (3) realising this inflection as the suffix `-es`. Similar, for `subject` the system has to infer that it should be inflected as plural number and realise it as the suffix `-s`. Note that we only focus on derivations, therefore, `proudly`, although being relevant to syntax, is not affected in this case. Taking it from the other perspective, this is also similar to the problem of generating text from tectogrammatical structure that has been studied for a long time (Hajic et al., 2002; Ptáček and Žabokrtský, 2006). There, the model

is provided with a dependency tree of content words. Our task differs in the way that we: (1) provide the tokens in their original order; and (2) use lemmatisation, therefore the categories that are realised non-morphologically are left unaffected. More recent work has also focused on abstract meaning representation (AMR) to text generation (Song et al., 2017).

Past work in supervised computational morphology—including the recent SIGMORPHON shared tasks on morphological reinflection (Cotterell et al., 2016a, 2017a) that was discussed in Section 2.3.3.2—has focused mainly on step (3) above. There, most neural models achieve high accuracy on many languages at type-level prediction of the form from its lemma and slot.

Our task amounts to a highly constrained version of language modelling. Language modelling predicts all words of a sentence from scratch, so the usual training and evaluation metric—perplexity—is dominated by the language model's ability to predict *content*, which is where most of the uncertainty lies. Our task focuses on just the ability to reconstruct certain missing parts of the sentence—inflectional morphemes and their orthographic realisation. This refocuses the modelling effort from semantic coherence to morphosyntactic coherence, an aspect of language that may take a back seat in current language models (see Belinkov et al., 2017; Linzen et al., 2016). Our task loosely resembles the C-test that is widely used to assess the competence of human second-language learners (Eckes and Grotjahn, 2006; Klein-Braley and Raatz, 1984) by requiring them to fill in the missing second halves of selected words throughout the occurring text.

Contextual inflection does not perfectly separate grammaticality modelling from content modelling. Mapping `Moira teach the student to write` to *`Moira taught the students to write`* does not require full knowledge of English grammar—the system does not have to predict the required word order nor the required infinitive marker `to`, as these are supplied in the input. Conversely, this example does still need to predict some content—the semantic choices of past tense and plural object are *not* given by the input and must be predicted by the system. A truer measure of grammatical competence would be a task of mapping a meaning representation to text, where the meaning representation specifies all necessary semantic content—content lemmata, dependency relations, and "inherent"

closed-class morphemes (semantic features such as noun number, noun definiteness, and verb tense)—and the system is to realise this content according to the morphosyntactic conventions of a language, which means choosing word order, agreement morphemes, function words, and the surface forms of all words.

Although our task is not perfectly matched to grammaticality modelling, the upside is that it is a "lightweight" task that works directly on text. No meaning representation is required. Thus, training and test data in any language can be prepared simply by lemmatising a naturally occurring corpus. Models for our task could be used to help detect and fix grammar errors in text for which no meaning representation is available, such as student writing or the output of neural machine translation.

Although few resources are required to construct the training and test data, additional annotated resources may still help to build a better system. Here we construct a system that is trained using fine-grained morphological tags—as well as a system that predicts forms directly without using any such morphological resources. We evaluate trained systems on 18 languages; examples from Polish are given in Table 4.1.

4.2 Predicting Inflectional Morphology

4.2.1 Task Notation

Given a language, let $\mathscr{M}$ be a language-specific set of morphological tags. Each $m \in \mathscr{M}$ has the form $m = \langle t, \sigma \rangle$, where t is a part of speech, and the slot σ is a set of attribute-value pairs that represent morphosyntactic information, such as those discussed in Section 4.2.2 below. We take $t \in \mathscr{T}$, the set of universal parts of speech described by Petrov et al. (2012).

Let Σ be the set of orthographic characters in the language. A word form $w \in \Sigma^+$ is a string of characters. When discussing the individual orthographic characters in a word w, we will also refer to the word by its sequence of such characters $\mathbf{c} = c_1 \cdots c_{|w|}$.

A sentence consists of a finite word sequence $\mathbf{w}$ (we use boldface for sequence variables). For every word w_i in the sequence, there is a corresponding analysis in terms of a morphological tag $m_i \in \mathscr{M}$ and a lemma ℓ_i. A lemma is itself a word—essentially a version of w with

(1)	*Jenia*	*daje*	*Maszy*	*ciekawą*	*książkę*
	John.M.SG.NOM	give.PRES.3SG	Mary.F.SG.DAT	interesting.F.SG.ACC	book.F.SG.ACC
(2)	*Książkę*	*ciekawą*	*Jenia*	*Maszy*	*daje*
	Book.F.SG.ACC	interesting.F.SG.ACC	John.M.SG.NOM	Mary.F.SG.DAT	give.PRES.3SG
(3)	*Jenia*	*Maszy*	*daje*	*ciekawą*	*książkę*
	John.M.SG.NOM	Mary.F.SG.DAT	give.PRES.3SG	interesting.F.SG.ACC	book.F.SG.ACC
(4)	*Jenie*	*Masza*	*daje*	*ciekawą*	*książkę*
	John.M.SG.DAT	Mary.F.SG.NOM	give.PRES.3SG	interesting.F.SG.ACC	book.F.SG.ACC

Table 4.1 Example sentences in Polish. While sentences (1), (2), and (3) use different word order, they are semantically equivalent and thus use the same inflected words. Given the lemma sequence from sentence (3), our system would have to guess semantically that it is inflected as (3), rather than as (4), which has a different meaning.

minimal inflectional marking. For example, the lemma of an English verb is conventionally taken to be its bare infinitive form, so `teach`, `teaches`, and `taught` all share the lemma `teach`. In general, w_i is determined by the pair $\langle \ell_i, m_i \rangle$. Although w_i can sometimes be computed by concatenating ℓ_i with m_i-specific affixes, it may sometimes be irregular. We adopt a "word-based morphology" approach (Aronoff, 1976; Spencer, 1991) that does not make assumptions about the relationship between w_i and $\langle \ell_i, m_i \rangle$ and, therefore, does not need any explicit decomposition of the forms; in our experiments, we will use a recurrent neural network to model this relationship.

Our proposed task is to predict a sentence $\mathbf{w}$ from its lemmatised form $\boldsymbol{\ell}$, sometimes inferring $\mathbf{m}$ as an intermediate latent variable. Our dataset (Section 4.5.1) provides all three sequences for each sentence.

4.2.2 Morphological Attributes

The following morphological attributes are of particular interest for our task and will figure into our error analysis later. They are encoded in our dataset (Section 4.5.1) according to the Universal Dependencies scheme. The attributes below are common across languages and tend to be realised by inflectional morphemes. We omit attributes such as mood and voice, as these are typically expressed instead by separate words or syntactic configurations.

- **Aspect** (Binnick, 2012) describes whether an action or event is continuing or has finished. In many Germanic languages it is combined with the tense category, while in Slavic, Basque, and Latin it is more autonomous. Unlike in Slavic, where aspect is typically expressed by derivation from the base (imperfective) form, Basque and Latin use inflection.
- **Case** (Blake, 2001; Butt, 2006; Malchukov and Spencer, 2009) marks a grammatical function, or role (subject, object, recipient, destination, possession, etc.) of a noun, pronoun, adjective, numeral and participle within a clause or a sentence. English presents a relatively simple case system, where only pronouns are marked morphologically. In this respect, the Uralic, Slavic and Romance languages are richer in their declension system.
- **Degree** is a feature specific to adjectives and adverbs. Mostly it is expressed either morphologically (as in English comparative `-er` or syntactically (as in English comparative `more`. Some languages such as Finnish or Basque mainly mark it morphologically and the same syntactic frames apply to all degrees, making the category harder to predict.
- **Definiteness** (Lyons, 1999) is presented in English, although it is realised outside of the word's morphology. This is the case in many Indo-European ("IE") languages. But in Bulgarian, it is realised as a definite article postfixed to nouns and adjectives, making its prediction more problematic. In Swedish, noun definiteness is expressed by means of articles and suffixes, and highly affects their declension.
- **Gender** (Corbett, 1991) initiates from nouns, and participates in their agreement with adjectives, verbs, articles and pronouns. Most IE languages exhibit gender to some extent, as either a binary or ternary system. Some Germanic languages, such as Danish, do not differentiate masculine and feminine. Basque only presents animacy, but not gender differences.
- **Number** (Corbett, 2000) is one of the most common grammar features, and present in most languages. Usually a language distinguishes either singular–plural or singular–dual–plural.
- **Person** (Siewierska, 2004) refers to the action participants. It is explicitly marked on pronouns in most languages, while verbs and auxiliaries require agreement in this category.
- **Tense** (Comrie, 1985) provides information about the time of the event in regards to the moment of speaking. Usually past and present forms are more likely to be expressed

morphologically, whereas the future form is often realised with auxiliaries. Tense categories are closely related to aspect, and in some languages, such as Basque, only aspect is specified.

Note that some of the above-mentioned features may be expressed through word order or periphrastic constructions in certain languages. For instance, English (and most other Germanic languages) mark the future tense through a periphastic construction: *Moira will proudly teach two subjects.* Spanish, on the other hand, would inflect the verb itself.[1]

4.3 An Encoder–Decoder Model

First we propose an encoder–decoder model that takes sentential (lemmata) context $\boldsymbol{\ell}$, combines it with a target form lemma representation ℓ_i and then decodes it into a word form conditioned on predicted history $w_{<i}$, i.e. without morphological feature prediction:

$$p(\mathbf{w} \mid \boldsymbol{\ell}) = \left(\prod_{i=1}^{n} p(w_i \mid \ell_i, w_{<i}, \boldsymbol{\ell}) \right) \tag{4.1}$$

The encoder performs an affine transformation on a concatenation of the lemmata together with predicted contexts and a character-level lemma form representations obtained as a last hidden states of corresponding BiLSTMs (Hochreiter and Schmidhuber, 1997).

The resulting representation $\mathbf{o}_i$ is fed into the decoder that produces the inflected target form character-by-character:

$$p(c_j \mid \mathbf{c}_{<j}, \mathbf{o}_i) = \text{softmax}\left(R \cdot \mathbf{c}_{<j} + \max\left(B \cdot \mathbf{o}_i, S \cdot \ell_j\right) + \mathbf{b}\right) \tag{4.2}$$

so the conditionals are then multiplied in order to get the sequence probability:

$$p(\mathbf{c} \mid \mathbf{o}_i) = \prod_{j=1}^{|w_i|} p(c_j \mid \mathbf{c}_{<j}, \mathbf{o}_i) \tag{4.3}$$

[1] This means that a subset of the above features will be captured in the reinflection task, depending on the language.

where c_j is the j-th character of the infected form, ℓ_j is the corresponding lemma character, B, S, R are weight matrices, and $\mathbf{b}$ is a bias term.

We now elaborate on the design choices behind the model architecture which have been tailored to our task. We supply the model with the ℓ_j character prefix of the lemma form to enable a copying mechanism, to bias the model to generate an inflected form that is morphologically-related to the lemma. In many cases, the inflected form is longer than its stem, and accordingly, when we reach the end of the lemma form, we continue to input an end-of-word symbol. We provide the model with the context vector $\mathbf{o}$ at each decoding step. It has been previously shown (Hoang et al., 2016) that this yields better results than other means of incorporation.[2] Finally, we use max pooling to enable the model to switch between copying of a lemma or producing a new character.

4.4 A Structured Neural Model[3]

We now introduce a more sophisticated model that also relies on morphological tags. In order to generate sequences of inflected forms from lemmata, we define the probability model:

$$p(\mathbf{w}, \mathbf{m} \mid \boldsymbol{\ell}) = \left(\prod_{i=1}^{n} p(w_i \mid \ell_i, m_i) \right) p(\mathbf{m} \mid \boldsymbol{\ell}) \tag{4.4}$$

In other words, the distribution is over interleaved sequences of one-to-one aligned inflected words and morphological tags, conditioned on a sequence of lemmata—all of length n. This distribution is drawn as a hybrid (directed-undirected) graphical model (Koller and Friedman, 2009) as illustrated on Figure 4.2. We define the two conditional distributions in the model in Section 4.4.1 and Section 4.4.2, respectively. As argued above, this model serves as a useful tool for studying natural language generation in morphologically complex languages.

[2] We tried to feed the context information at the initial step only, and this led to worse prediction in terms of context-aware suffixes.

[3] The model was developed in collaboration with Jason Eisner and Ryan Cotterell

4.4.1 A Neural Conditional Random Field

The distribution $p(\mathbf{m} \mid \boldsymbol{\ell})$ is defined to be a conditional random field (CRF). CRFs were first introduced by Lafferty et al. (2001) as a generalisation of classical maximum entropy models (Berger et al., 1996) to globally normalised distributions over structured objects. In this work, our CRF is a conditional distribution over morphological taggings of a sequence of lemmata. We define this conditional distribution as

$$p(\mathbf{m} \mid \boldsymbol{\ell}) = \frac{1}{Z(\boldsymbol{\ell})} \prod_{i=1}^{n} \psi(m_i, m_{i-1}, \boldsymbol{\ell}),$$

where $\psi(\cdot,\cdot,\cdot) \geq 0$ is an arbitrary potential[4] and $Z(\boldsymbol{\ell})$ normalises the distribution. Recall that $Z(\boldsymbol{\ell})$ may be computed in linear time with the forward algorithm.

In this work, we opt for a recurrent neural potential. Specifically, we adopt a parameterisation similar to the one given in Lample et al. (2016), but we remark that neural CRFs have a much longer history in the literature (Artieres et al., 2010; Collobert et al., 2011; Fujii et al., 2012; Peng et al., 2009; Vinel et al., 2011; Wang and Manning, 2013). Our potential ψ is computed as follows. First, the sentence of lemmata is encoded into a sequence of word vectors using the strategy described by Ling et al. (2015a): a unidirectional LSTM is folded over the character sequence (each character is encoded as a one-hot vector) and the final hidden state of this LSTM is taken as the word vector. Then, these character-infused word vectors are passed to a bidirectional LSTM (Graves et al., 2005), which consists of two unidirectional LSTMs, one run left-to-right and the other right-to-left, with the corresponding hidden states concatenated at each time step. Notationally, we will simply refer to the hidden state $\mathbf{h}_i \in \mathbb{R}^d$ as the result of said concatenation at the i-th step. Using $\mathbf{h}_i$, we can define the potential function as

$$\psi(m_i, m_{i-1}) = \exp\left(a_{m_i, m_{i-1}} + \mathbf{o}_{m_i}^{\top} \mathbf{h}_i\right), \tag{4.5}$$

where $a_{m_i, m_{i-1}}$ is a transition weight, identical to those found in linear-chain CRFs and $\mathbf{o}_{m_i} \in \mathbb{R}^d$ is a morphological tag embedding; both to be learned.

[4] We slightly abuse notation and use m_0 as a distinguished beginning-of-sentence symbol.

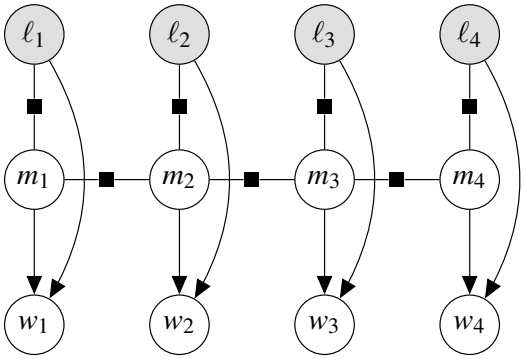


Figure 4.1 Our structured neural model shown as a hybrid (directed-undirected) graphical model.

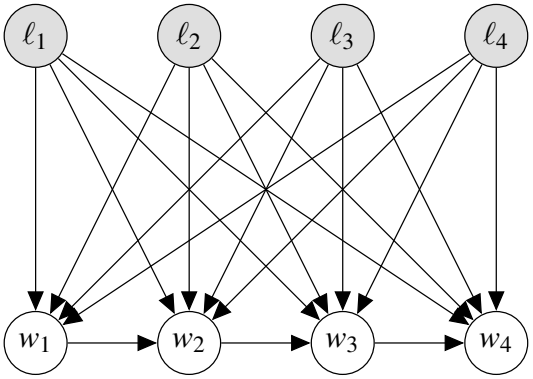


Figure 4.2 Neural encoder-decoder model shown as a graphical model.

4.4.2 The Morphological Inflector

The conditional distribution $p(w_i \mid \ell_i, m_i)$ is parameterised by a neural sequence-to-sequence model (Sutskever et al., 2014) with attention (Bahdanau et al., 2015). The model is an encoder-decoder, where the input sequence is encoded into a sequence of fixed-length vectors using a bidirectional LSTM, and then the output sequence is decoded character-by-character. Suppose we denote the k-th hidden state as output of our bidirectional LSTM, with $\mathbf{h}_k^{(enc)}$. Bahdanau attention dictates we take a convex combination of all the hidden states $\mathbf{a}_{j-1} = \sum_{k=1}^{|w_i|} \alpha_k(j-1) \cdot \mathbf{h}_k^{(enc)}$, where the $\alpha_k(j-1)$ themselves are determined by a multi-layer perceptron; we predict the j-th character from

$$p(c_j \mid \mathbf{c}_{<j}, m_i) = \text{softmax}\left(\mathbf{W} \cdot \mathbf{a}_{j-1} + \mathbf{b}\right) \tag{4.6}$$

where, recall, $w_i = \mathbf{c} = c_1 \cdots c_{|w_i|}$. The distribution in Eq. (4.6), strung together with the other conditionals, yields a joint distribution over the entire character sequence:

$$p(\mathbf{c} \mid \ell_i, m_i) = \prod_{j=1}^{|w_i|} p(c_j \mid \mathbf{c}_{<j}, m_i) \tag{4.7}$$

To condition the model on m_i, we adopt the encoding of Kann and Schütze (2016), who input the actual character string $\mathbf{c}$ prepended with its morphological tag into the network. For instance, to map the lemma *talk* to its gerund *talking*, we feed in `<w> V GERUND t a l k </w>` and train the network to output `<w> t a l k i n g </w>`, where we have augmented the orthographic character alphabet Σ with the feature-attribute pairs that constitute the morphological tag m_i.

4.4.3 Parameter Estimation and Decoding

We optimise the log-likelihood of the training data with respect to all model parameters. As Eq. (4.4) is differentiable, this may be achieved with standard gradient-based methods, such as backpropagation (Rumelhart et al., 1986). The exact details are found in Section 4.5.3. Note that since we estimate the parameters in the fully supervised case, the directedness

Language	UD	Family	Type	Order	Slots	Tokens
Basque	eu	Isolated	$\mathscr{A}$	SOV	845	73k
Bulgarian	bg	Slavic (IE)	$\mathscr{F}$	Free	447	125k
Czech	cs	Slavic (IE)	$\mathscr{F}$	Free	1552	183k
Danish	da	Germanic (IE)	$\mathscr{F}$	SVO	160	89k
Dutch	nl	Germanic (IE)	$\mathscr{F}$	SVO	325	188k
English	en	Germanic (IE)	$\mathscr{F}$	SVO	118	205k
Finnish	fi	Uralic	$\mathscr{A}$	SVO	1310	127k
Hindi	hi	Indic (IE)	$\mathscr{F}$	SOV	921	281k
Hungarian	hu	Uralic	$\mathscr{A}$	Free	424	21k
Irish	ga	Celtic (IE)	$\mathscr{F}$	VSO	371	17k
Italian	it	Romance (IE)	$\mathscr{F}$	SVO	269	249k
Latin	la	Romance (IE)	$\mathscr{F}$	Free	910	246k
Norwegian	no	Germanic (IE)	$\mathscr{F}$	SVO	169	245k
Polish	pl	Slavic (IE)	$\mathscr{F}$	Free	613	70k
Portuguese	pt	Romance (IE)	$\mathscr{F}$	Free	492	202k
Spanish	es	Romance (IE)	$\mathscr{F}$	SVO	389	382k
Slovenian	sl	Slavic (IE)	$\mathscr{F}$	Free	1180	112k
Swedish	sv	Germanic (IE)	$\mathscr{F}$	SVO	131	67k

Table 4.2 A list of languages used for the experiments. Here $\mathscr{F}$ and $\mathscr{A}$ stand for fusional and agglutinative language, respectively. Also note that for Slavic languages SVO order is more natural and used more often than others.

allows independent estimation of the CRF (described in Section 4.4.1) and the inflector (described in Section 4.4.2) independently.

Decoding, on the other hand, is a bit more complicated. We opt for a greedy strategy where we first decode the CRF, that is, we solve the problem

$$\mathbf{m}^\star = \operatorname*{argmax}_{\mathbf{m}} \log p(\mathbf{m} \mid \boldsymbol{\ell}), \tag{4.8}$$

for which Viterbi (1967) provides us with a linear-time (in $|\boldsymbol{\ell}|$) exact algorithm. We then use this decoded $\mathbf{m}^\star$ to generate forms from the inflector using beam search, as is common in other generation tasks performed with a neural sequence-to-sequence model, such as machine translation.

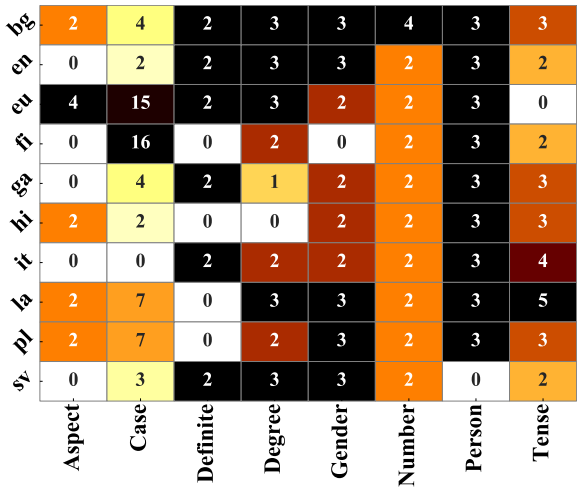


Figure 4.3 Number of possible values per morphological attributes for each language; the darker colors correspond to more possible values. The numbers indicate the maximum values for an attribute in the UD schema.

4.5 Experiments

4.5.1 Dataset

We use the Universal Dependencies v1.2 dataset (Nivre et al., 2016) for our experiments. We include all the languages with information on their lemmata and fine-grained grammar tag annotation that also have `fasttext` embeddings described in Section 2.3.5, which are used for word embedding initialisation.[5] Table 4.2 lists the languages along with their corresponding sizes of the training data and their slot sets. Figures 4.3 and 4.4 illustrate how many values each morphological attribute can take, ranging from 118 to 1552. Importantly, although being almost complete, it is not an exhaustive list of unseen slots.

We additionally note that some inconsistency and ambiguity still remain in the data in cases of syncretism. For instance, there is a difference in annotation when multiple genders or cases share the same form: some languages specify all the possibilities whereas others provide a single one.

[5] We also choose mainly non-Wikipedia datasets to reduce any possible intersection with the data used for the *FastText* model training

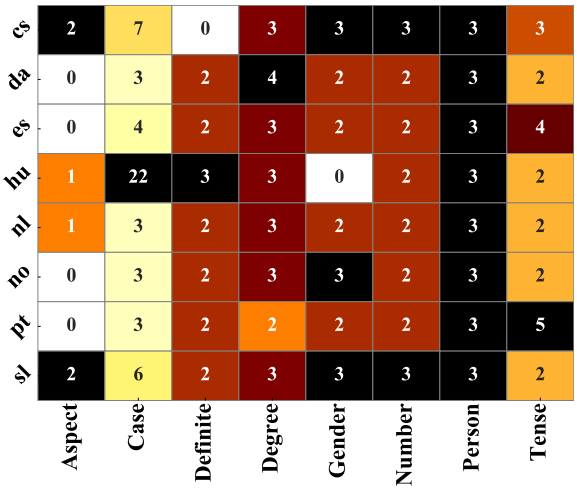


Figure 4.4 Related Languages: number of possible values per morphological attribute for each language; the darker colors correspond to more possible values in each column.

4.5.2 Evaluation

We evaluate our model's ability to predict: (i) the correct morphological tags from the lemma context, and (ii) the correct inflected forms. As our evaluation metrics, we report 1-best accuracy for both the tags and form prediction.

Skyline: The Morphological Cloze. Here we present a skyline: a point of comparison we expect to lose out to. Many of the tags and forms we wish to predict from the sequence of lemmata are not easily guessed without additional information, as discussed in Section 4.1. To compare our model against a second system, we consider a related task, which we term the **morphological cloze** task. Here, we use the same model as described in Eq. (4.4), but provide a sentential *gold* context. That is, we give the model the actual observed forms except for the lemma, whose tag and inflection we seek to predict. Naturally, the closer our model is to this skyline, the better it does.

Direct Form Prediction Baseline with Neural Model. As a baseline for the form prediction, we additionally adopt a more lightweight setting that does not rely on morphological tags. More specifically, we train a neural encoder-decoder model as described in Section 4.3 to predict inflected forms directly from a sequence of lemmata (the contextual inflection task)

or an inflected context (the skyline). Note that the model is similar to the morphological inflector described above with two essential differences: (i) we use sentence-level contextual representation instead of explicit morphological tagset m_i; and (ii) we do not use an attention mechanism.

Agreement Evaluation We consider the relations that typically require agreement, such as: (1) verb-subject (noun and pronoun) *nsubj*; and (2) adjective-noun *amod* and do not evaluate polypersonal agreement. Here, we merely expect a match in a corresponding morphological category in both predicted forms if there is such a match in their initial annotation. We empirically evaluate the categories that should be in agreement from the training data. We only consider the cases when the part of speech is guessed correctly (which is mostly predicted at 95–98% accuracy).

4.5.3 Hyperparameters and Other Minutiae

We use a word and character embedding dimensionality of 300 and 100, respectively. The hidden state dimensionality is set to 100 and 200 for the both encoder–decoders and CRF, respectively. For both models we choose LSTMs as RNN units. Models are trained with Adam (Kingma and Ba, 2014) with a learning rate of 0.001 for 20 epochs.

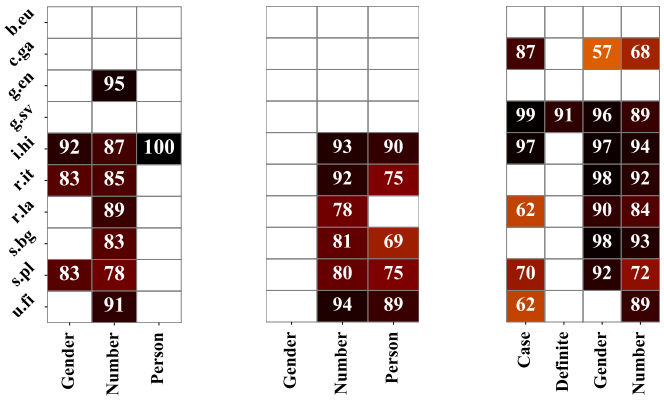
4.6 Results, Error Analysis, and Discussion

Table 4.3 presents the accuracy of our best model across all languages. Given that our task is novel, we cannot readily benchmark against the work of others. So, what has contextual inflection taught us about natural language generation, especially in morphologically rich languages? We highlight five lessons from our error analysis that apply to a wider range of generation tasks, e.g., machine translation and dialogue systems.

① **Neural Networks Learn Agreement.** In all the languages under consideration, the inflection of the adjective with respect to its gender depends on the noun. Noun gender

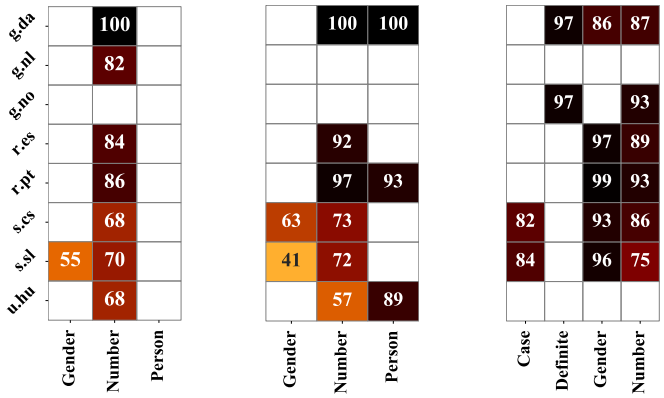
$\mathscr{L}$	Family	tag (s)	tag (c)	form (g)	form (s)	form (c)	form (direct)
eu	Isolated	66.63$^{4.19}$	64.97$^{4.77}$	82.19	61.05	59.91	58.28
ga	Celtic	68.33$^{4.16}$	66.65$^{4.68}$	84.50	69.53	67.78	64.48
da	Germanic	86.26$^{1.67}$	82.07$^{1.80}$	97.31	87.16	83.15	79.46
en	Germanic	89.58$^{1.50}$	88.38$^{1.53}$	95.57	90.41	89.49	86.75
nl	Germanic	82.70$^{2.30}$	80.05$^{2.96}$	88.29	81.30	79.17	76.68
no	Germanic	87.16$^{1.66}$	83.48$^{1.71}$	91.59	82.40	78.70	83.10
sv	Germanic	81.86$^{1.99}$	76.35$^{2.45}$	96.02	82.47	77.61	74.75
hi	Indic	85.33$^{1.82}$	82.39$^{2.00}$	87.49	81.43	79.73	89.71
es	Romance	85.89$^{1.81}$	78.59$^{2.22}$	95.17	87.95	81.07	78.49
it	Romance	92.28$^{1.39}$	84.48$^{1.68}$	85.13	80.39	73.94	82.49
la	Romance	82.57$^{2.19}$	71.65$^{2.95}$	89.69	75.68	66.31	68.39
pt	Romance	88.22$^{1.58}$	80.75$^{1.90}$	98.21	91.25	84.25	82.75
bg	Slavic	81.55$^{2.07}$	76.07$^{2.38}$	91.89	78.81	73.75	71.45
cs	Slavic	76.39$^{3.30}$	68.07$^{4.23}$	97.38	80.56	73.04	64.12
pl	Slavic	71.94$^{3.34}$	63.25$^{4.17}$	96.14	74.83	66.65	58.90
sl	Slavic	78.82$^{2.38}$	67.79$^{3.21}$	97.71	81.79	71.10	62.78
hu	Uralic	68.22$^{4.57}$	67.74$^{5.07}$	86.31	62.45	61.35	68.29
fi	Uralic	65.99$^{5.10}$	58.20$^{6.57}$	86.53	59.34	52.05	49.99

Table 4.3 Accuracy of the models for various prediction settings. The column header **tag** refers to the tag prediction accuracy and **form** refers to the form prediction accuracy. We mark the contextual inflection setting with (c) and the skyline with (s). We additionally report accuracy achieved in form prediction only from the gold tags (g). Blue superscripts correspond to perplexity values (lower is better).



(a) Noun-Verb (b) Pron-Verb (c) Adj-Noun

Figure 4.5 Contextual inflection agreement results. The *x*-axis shows three morphological attributes: gender, number and person. The *y*-axis shows the language names.

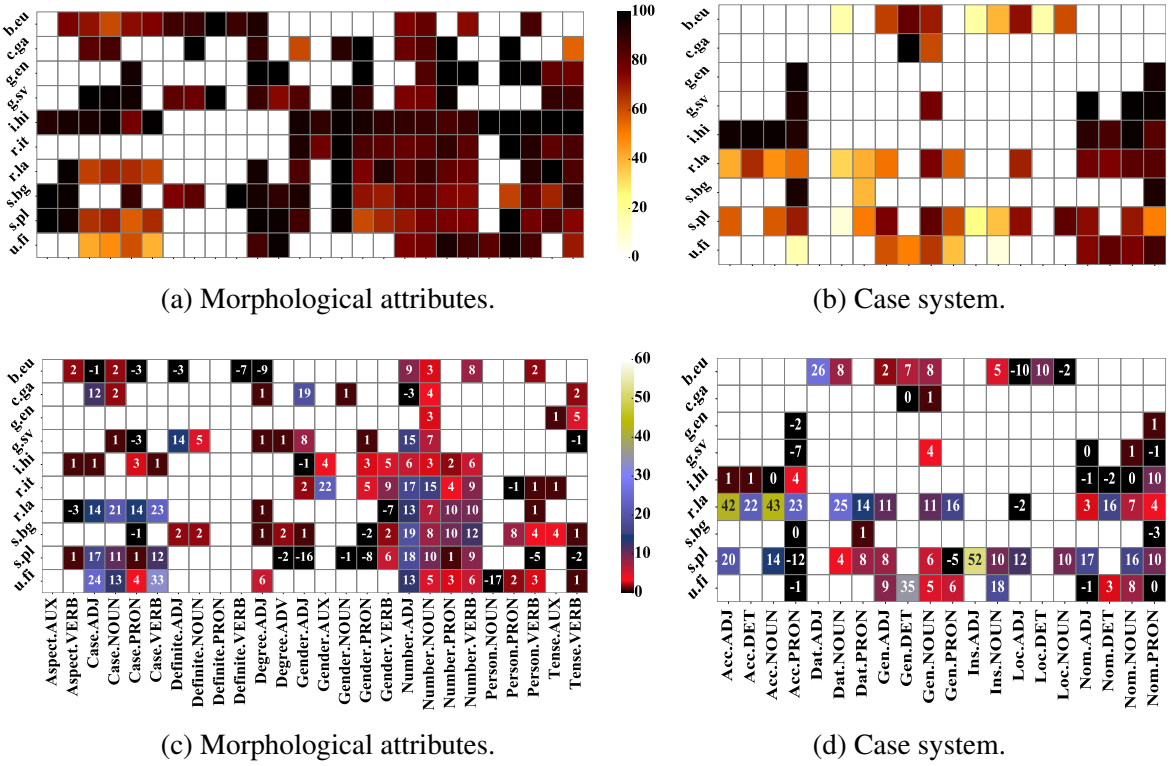


(a) Noun-Verb (b) Pron-Verb (c) Adj-Noun

Figure 4.6 Related Languages: contextual inflection agreement results for subject-verb (*nsubj*) and adjective-noun (*amod*). The *x*-axis shows morphological attributes. The *y*-axis shows the language names.

is often marked on the lemma itself, and, as Table 4.7a shows, the network easily learns the proper concord—properly declining adjectives to match the gender of the head. Verbal gender, which appears in the past tense of many Slavic languages, seems to be harder to predict. Given that the linear distance between the subject and the verb may be longer, we suspect the network struggles to learn the longer-distance dependencies, consistent with the findings of Linzen et al. (2016). Like noun gender, pronoun person is not lost during lemmatisation and our networks, likewise, achieve high accuracy in predicting the proper concord with the verb on most languages with the exception of Bulgarian. Interestingly, we note that performance of the purely lemma-based prediction task is similar to the cloze skyline: see Table 4.7c. Now, we turn to number, which is often expressed morphologically, and, which we find is harder to predict. Often, the context is simply not enough to tell us what number we need to predict. For this reason, Table 4.7c shows that we are often 10–20 points lower than the skyline for number prediction.

② **Morphological Complexity Matters.** In the previous paragraph, we observed that our model learns gender agreement quite well in many cases—matching the skyline at times. However, there is a notable exception: for languages with rich case systems, e.g., the Slavic languages (which exhibit a lot of fusion) and the agglutinative Uralic languages, performance



(a) Morphological attributes.

(b) Case system.

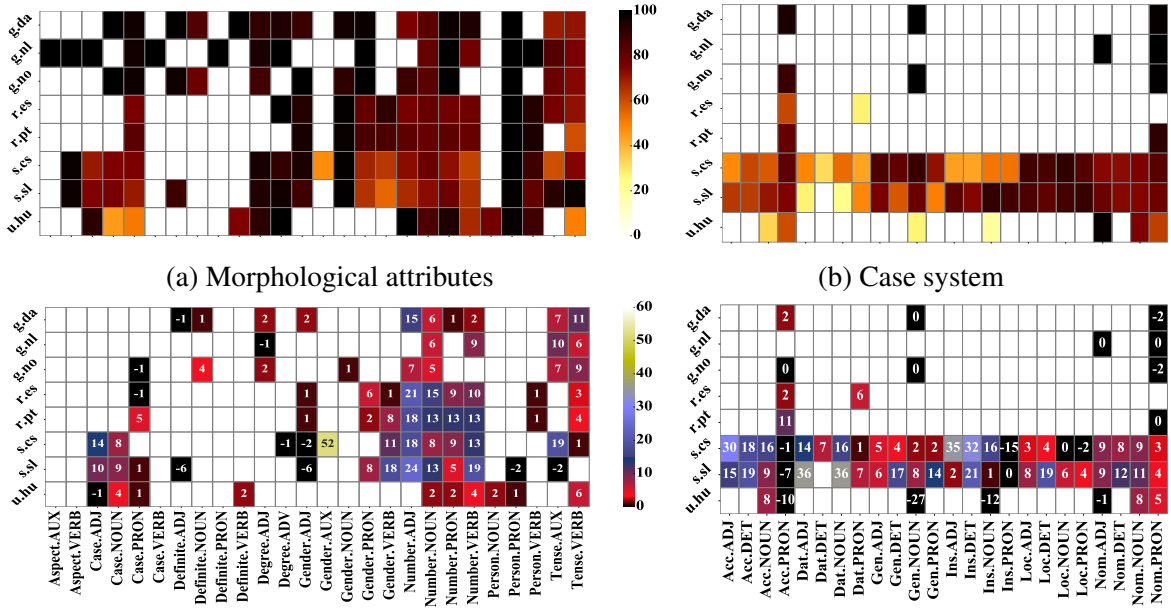
(c) Morphological attributes.

(d) Case system.

Figure 4.7 We display heatmaps that show how often morphological value was correctly predicted. The *x*-axis is the morphological value and the *y*-axis is the language name. The top two figures provide tag prediction results (from lemmata) organised by language family. The bottom two figures show the differences in morphological tag prediction between contextual inflection and the skyline. In the bottom two figures, positive numbers indicate how many points the skyline wins by.

is much worse, as evidenced in Table 4.5. This suggests that generation in languages with more morphological complexity will be a harder problem for NLP to solve. Indeed, this problem is under-explored, as the field of NLP tends to fixate on generating English text, e.g., in machine translation or dialogue system research. We suggest expanding the focus of generation research to morphologically rich languages.

③ **Predictability of Inherent Categories.** As Booij (1996) mentions, tense is an inherent category and, in the languages where it is mainly expressed morphologically, it becomes hard to predict correctly (unless there is no strong signal within a sentence such as occurs with *yesterday*, *tomorrow*, or *ago*). And, indeed, for most languages, with the exception of Hindi and Slovenian, it is still challenging to achieve good accuracy. On the other



(a) Morphological attributes

(b) Case system

(c) Morphological attributes

(d) Case system

Figure 4.8 Related Languages. We display heatmaps that show how often morphological value was correctly predicted. The *x*-axis is the morphological value and the *y*-axis is the language name. Note that the white cells mean the given language does not express that given morphological attribute. The top two figures provide tag prediction results (from lemmata) organised by language family. In the bottom two figures, positive numbers indicate how many points the skyline wins by.

hand, aspect, although being closely related to tense, is well-predicted since it is mainly expressed as a separate lexeme. But, in general, it is still problematic to make a prediction in languages where it is morphologically marked or highly mixed with tense, as in Basque. We suggest that correct prediction of tense will require document-level context, when translating from a language that does overtly mark it, e.g., Mandarin Chinese, or in dialogue systems. Future iterations of our model will exploit such document-level signals. Definiteness, another inherent category, is well predicted in the languages where it has non-morphological expression since it is not affected by lemmatisation. For example, if we look at Bulgarian, we see that it is, indeed, extremely problematic, because it is highly dependent on the speaker's intention.

④ **Prediction of Grammatical Case.** As illustrated in Figures 4.7b, 4.7d, 4.8b, and 4.8d, our model finds prediction of case to be extremely challenging in morphologically complex

languages such as Hungarian, Finnish or Basque. For the languages that allow free word order, we observe a substantial improvement in the skyline setting. A closer look shows that the models are able to predict the subject of the sentence, which is typically marked by nominative case, although the object, marked by accusative or dative, seems to be a lot harder to capture, especially in the contextual inflection setting. Interestingly, genitive case marking possession or absence (typically expressed by a modifier in English) yields high accuracy. Locative case is often accompanied by a post/preposition, providing a strong signal and leading to better prediction, close to the skyline.

⑤ **Use of Latent Morphology.** Finally, the comparison of our structured model with latent morphological tags (at test time) to the direct form generation baseline suggests that we should be including linguistically motivated latent variables into models of natural language generation. We observe in Table 4.3 that predicting the tag together with the form often improves performance. While some recent work in neural machine translation (Klein et al., 2017; Tamchyna et al., 2017) has made use of target-side morphology, this is still not standard.

4.7 SIGMORPHON 2018 – SubTask 2

The contextual inflection task described here motivated the SIGMORPHON 2018 shared task organisers to run a related sub-task (Cotterell et al., 2018). Unlike the task stated above, they did not aim to infect *all* lemmata in the sentence but rather provided inflected context and only predicted a few (1-3) forms.[6] The task comprised of two tracks. In the first track lemmata and morphosyntactic descriptions of all contextual words are both additionally provided:
(1) `The/the+DT ___(dog) are/be+AUX;IND;PRS;FIN barking/bark+V;PTCP`,
and the systems were required to predict the target form. The second track only provides contextual forms and, therefore, is more challenging:
(2) `The ___(dog) are barking`.

[6]Therefore, the task is more similar to the Skyline setting.

In both tracks systems need to predict `dogs`. Both tracks are run in three different data settings depending on the number of tokens available to train on: high (10^5), medium (10^4), and low (10^3).

The data for the task was sampled from the Universal Dependencies v.2 treebanks (Nivre et al., 2017), and morphological annotations were converted into the UniMorph format. The submitted systems were evaluated in terms of their ability to predict: (1) the original word form; and (2) a contextually plausible word form even if it's different from the original (as in `We ____(see) the dog.` where both `see` and `saw` fit the context). The latter one requires each test sample to be annotated manually (whether it is grammatically correct), and, therefore, limits the number of languages for this task. In total, the task covers seven languages: English, Finnish, French, German, Russian, Spanish, and Swedish. The sentences that contained a token found in UniMorph (i.e. its lemma and morphological tags presented there) were sampled from the UD dataset.

Submitted systems were evaluated based on accuracy and average Levenshtein distance between the prediction and the truth. The baseline was inspired by an encoder–decoder with attention mechanism used for the re-inflection task (Kann and Schütze, 2016). More specifically, it is conditioned on left and and right context words augmented by left and right context lemmas and a character-level representation of the target lemma. The decoder using an attention mechanism generates the output form character-by-character.[7] The second baseline system just copied lemmas to the output.

All teams submitted neural systems, of which all but one (which was a neural transition-based transducer with a copy action) were derived from Kann and Schütze (2016). This task appeared to be more challenging than the re-inflection one. The "copy" baseline achieved on average 36.62% accuracy for original and 42% for plausible forms. The neural baseline got 62.41% accuracy in predicting original forms in the high-resource setting of track 1 and 1.85% in low-resource. For plausible forms it received 69.53% and 2.63%, respectively. In the high-resource setting of track 2 the accuracy of the neural system dropped to 54.48% for original and 60.79% for plausible forms, while in the low-resource setting

[7]The hyperparameters were set as follows: all dimensionality was set to 100 for both encoder and decoder, the number of layers set to two. The system was trained for 20 epochs with Adam (Kingma and Ba, 2014).

it was 2.19% and 3.11%, respectively. The best performer for track 1 (Kementchedjhieva et al., 2018) outperformed the neural baseline by 6% in high-resource original and plausible form prediction. The system: (1) used an RNN over a sentence to predict morphological tags; and (2) combined all training data for all languages available for this task (therefore, it can be seen as multilingual). The system from Makarov and Clematide (2018) was the best performer in all low-resource settings. For track 1 it achieved 42.42% and 48.49% in original and plausible forms, respectively. In particular, Makarov and Clematide (2018) implemented a neural transition-based transducer enriched with a copy mechanism that was applied to transform a lemma to an output form. In addition, it also used beam search for decoding. The best performer in the track 1 high-resource setting also got highest accuracy in track 2 in the original form prediction setting, but the value dropped by 13% compared to track 1. In the low-resource setting it reduced to 38.60%. In the high-resource plausible form prediction setting none of the submitted systems outperformed the neural baseline; for the low-resource setting, the best accuracy dropped by 3 points.

4.8 Conclusion

We introduced the novel task of contextual inflection, whereby a sequence of correctly inflected forms is to be generated from a sequence of uninflected lemmata—we treat the morphological tag as a latent variable. Our goal is to provide a more intrinsic analysis of contemporary neural models and their ability to generate correctly inflected text. We developed a hybrid graphical model with a recurrent neural parameterisation and evaluated it on 18 languages. Our analysis showed that some morphological tags could be easily predicted from sentential context, as they participate in agreement. Others such as noun gender or verbal aspect are typically inherent, but still captured well based on more global sentential context. While the task is self-contained, we have highlighted several key points that will be crucial for many natural language generation systems in the years to come, especially when NLP practitioners attempt to develop systems for the generation of text in morphologically complex languages. In addition, we organised a shared task on contextual

inflection where the systems competed in the setting similar to the cloze task. We evaluated the performance of the systems in various settings by calibrating the amount of data available for training (low, medium, and high) and providing them with morphosyntactic tags. The results show that there is a significant gap between accuracies obtained with and without morphosyntactic annotation. We also observe a performance drop when the amount of data is reduced. This suggests that there is still a lot of room for improvement in the models' ability for generalisation.

Chapter 5

Derivational Morphology Models

A large part of the chapter appears in the following paper:

Ekaterina Vylomova, Ryan Cotterell, Timothy Baldwin, and Trevor Cohn. Context-aware prediction of derivational word-forms. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: (Volume 2 : Short Papers)*, pages 118–124, 2017b.

5.1 Introduction

In chapters 3 and 4 we mainly focused on studying of morphological inflection and showed that regularities existing in word changing processes are captured quite well by contemporary neural models. In this chapter, we turn to analysis of derivational processes. Understanding how new words are formed is a fundamental task in linguistics and language modelling, with significant implications for tasks with a generation component, such as abstractive summarisation and machine translation. In Section 3.2 we evaluated various types of lexical relations and showed that morphosyntatic relations are captured well compared to morphosemantic. In Section 2.3.3.2 we discussed a shared task on morphological reinflection. The results of the task demonstrated a superior performance of neural approaches and ability to achieve high accuracy even in low-resource conditions. Further, in Section 4.2

we introduced a new task of contextual inflection that aims to measure a model's ability to infer morphological features and generate inflected forms directly from sentential context. The experiments and results obtained in our experiments as well as SIGMORPHON 2018 sub-task 2 illustrated that the task is a lot more challenging than vanilla morphological reinflection with gold tags. This chapter aims at evaluation of derivational models on these two tasks.

Unlike in inflectional morphology modelling, there were quite a few works in the direction of computational derivational morphology, especially attempts at paradigmatic treatment of it. In Cotterell et al. (2017b) the authors[1] studied various English verbal nominalisations (AGENT (`employ` → `employer`), PATIENT (`employ` → `employee`), RESULT (`employ` → `employment`), adjective adverbalisations (`interesting` → `interestingly`) and nominalisations (`evident` → `evidence`)). They experimented with English derivational triples extracted from NomBank (Meyers et al., 2004) (e.g. `employ`+AGENT → `employment`) excluding zero-derivations (such as `rent` → `rent`), resulting in 6,029 derivational samples. The task was stated similarly to morphological inflection, i.e. given a base form together with target tags a model had to generate a derived form. The model for the form prediction was derived from the model used in the morphological infection task. In particular, the authors used the best performing model of the 2016 shared task on inflection described in Section 2.3.3.2, an encoder-decoder gated recurrent neural network (Bahdanau et al., 2015). The model was evaluated and compared against the same baseline system used for the reinflection task (see Section 2.3.3.2) in terms of accuracy and the Levenshtein distance between predicted and gold output strings. The neural encoder took a character-level representation of the base form and the target tag such as `a m e l i o r a t e RESULT` and then the decoder generated the target string `a m e l i o r a t i o n`. The results showed: 1) the superior performance of the neural model; 2) lower accuracy compared to inflectional morphology prediction; and 3) different levels of derivational slot regularity. Regarding the latter, adverbalisation appeared to be almost as productive and regular as inflection (it's mainly realised as the `-ly` suffix), it achieves 90%

[1] I am one of the co-authors, although, for the purpose of the thesis, this is not treated as a novel contribution.

accuracy in both neural and non-neural models (in 1-best prediction). The RESULT category, on the other hand, is more vague and less regular, leading to accuracy of 40% for FST-based and 53% for neural approaches (although raising to 70% for 10-best prediction). Agentives are a bit more regular leading to 52% and 65% (82% for 10-best), respectively. Indeed, the category does not differentiate between male and female agentives as well as `-er;-or` from `-ist`. An error analysis showed that many cases require additional information for prediction such as etymology (as in `*containee` and `content`), more specific meanings (as in the distinction between `complexity` and `complexness`), gender (as in `waiter` and `waitress`), and regularity (as in `*advancely` and `in-advance`).

To conclude, the authors showed that derivations are more challenging due to less regularity and more opaque meanings. At the same time, we can place derivations and inflections on a continuous scale of productivity and specificity, where inflections present more productive (and compositional) forms and meanings and can be applied to a wider range of lemmas. Derivations, on the other hand, often are less productive and more restrictive but still we can identify cases that behave almost like inflections such as adverbalisation by attaching a `-ly` suffix.

Similar to inflections, in the next section we study contextual prediction of derivations.

5.2 Context-Aware Prediction

In this part, we focus on modelling derivational morphology to learn, e.g., that the appropriate derivational form of the verb `succeed` is `succession` given the context *`As third in the line of ____...`*, but is `success` in *`The play was a great ____`*. Derivational paradigm completion task that was discussed earlier requires to specify paradigm slots, i.e. identify a paradigm structure. As mentioned in Section 2.2.5, derivations are less studied, present more problems in identifying regularities, and there is no agreement on paradigmatic treatment of derivational morphology. Therefore, here we consider replacing derivational slots with contextual representations. More specifically, we study predictability of derived forms from their sentential contexts.

As we discuss in Section 2.2.2, English is broadly considered to be a morphologically impoverished language, but there are certainly many regularities in morphological patterns, e.g., the common usage of `-able` to transform a verb into an adjective, or `-ly` to form an adverb from an adjective. However, there is considerable subtlety in English derivational morphology, in the form of: (a) idiosyncratic derivations; e.g. `picturesque` vs. `beautiful` vs. `splendid` as adjectival forms of the nouns `picture`, `beauty` and `splendour`, respectively; (b) derivational generation in context, which requires the automatic determination of the part-of-speech (POS) of the stem and the likely POS of the word in context, and POS-specific derivational rules; and (c) multiple derivational forms often exist for a given stem, and these must be selected based on the context (e.g. `success` and `succession` as nominal forms of `success`, as seen above). As such, there are many aspects that affect the choice of derivational transformation, including morphotactics, phonology, semantics or even etymological characteristics. Earlier works (Thorndike, 1941) analysed ambiguity of derivational suffixes themselves when the same suffix might present different semantics depending on the base form it is attached to (cf. `beautiful` vs. `cupful`). Furthermore, as Richardson (1977) previously noted, even words with quite similar semantics and orthography such as `horror` and `terror` might have non-overlapping patterns: although we observe regularity in some common forms, for example, `horrify` and `terrify`, and `horrible` and `terrible`, nothing tells us why we observe `terrorize` and no instances of `horrorize`, or `horrid` but not `terrid`.

In this part, we propose the new task of predicting a derived form from its context and a base form. Our motivation in this research is primarily linguistic, i.e. we measure the degree to which it is possible to predict particular derivation forms from context. A similar task has been proposed in the context of studying how children master derivations (Singson et al., 2000). In their work, children were asked to complete a sentence by choosing one of four possible derivations. Each derivation corresponded either to a noun, verb, adjective, or adverbial form. Singson et al. (2000) showed that childrens' ability to recognise the correct form correlates with their reading ability. This observation confirms an earlier idea that orthographical regularities provide a clearer clues to morphological transformations

compared to phonological rules (Moskowitz, 1973; Templeton, 1980), especially in languages such as English where grapheme-phoneme correspondences are opaque. For this reason we consider orthographic rather than phonological representations.

In our approach, we test how well models incorporating distributional semantics can capture derivational transformations. In this work, we deal with the formation of deverbal nouns, i.e., nouns that are formed from verbs. Common examples of this in English include agentives (e.g., `explain` ↦ `explainer`), gerunds (e.g., `explain` ↦ `explaining`), as well as other nominalisations (e.g., `explain` ↦ `explanation`). Nominalisations have varyingly different meanings from their base verbs, and a key focus of this study is the prediction of which form is most appropriate depending on the context, in terms of syntactic and semantic concordance. Our model is highly flexible and easily applicable to other related lexical problems.

5.2.1 Dataset

As the starting point for the construction of our dataset, we used the CELEX English dataset (Baayen et al., 1993). We extracted verb–noun lemma pairs from CELEX, covering 24 different nominalisational suffixes and 1,456 base lemmas. Suffixes only occurring in 5 or fewer lemma pairs mainly corresponded to loan words and consequently were filtered out. We augmented this dataset with verb–verb pairs, one for each verb present in the verb–noun pairs, to capture the case of a verbal form being appropriate for the given context.[2] For each noun and verb lemma, we generated all their inflections, and searched for sentential contexts of each inflected token in a pre-tokenised dump of English Wikipedia.[3] To dampen the effect of high-frequency words, we applied a heuristic log function threshold which is basically a weighted logarithm of the number of the contexts. The final dataset contains 3,079 unique lemma pairs represented in 107,041 contextual instances.[4]

[2] We also experimented without verb–verb pairs and didn't observe much difference in the results.

[3] Based on a 2008/03/12 dump. Sentences shorter than 3 words or longer than 50 words were removed from the dataset.

[4] The code and the dataset are available at `https://github.com/ivri/dmorph`

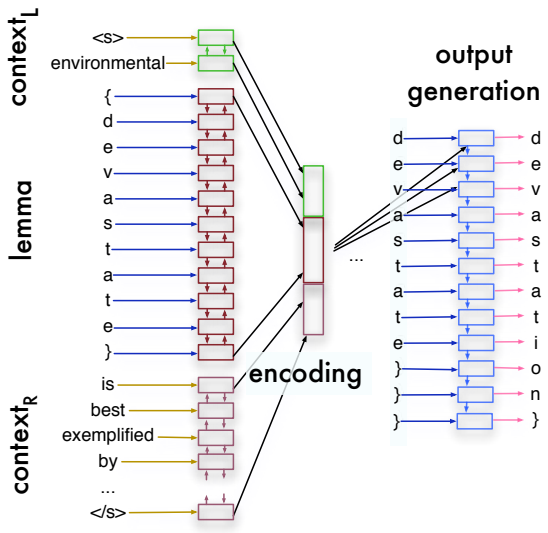


Figure 5.1 The encoder–decoder model, showing the stem `devastate` in context producing the form `devastation`. Coloured arrows indicate shared parameters.

5.2.2 Experiments

Here we model derivational morphology as a prediction task, formulated as follows. We take sentences containing a derivational form of a given lemma, then obscure the derivational form by replacing it with its base form lemma. The system must then predict the original (derivational) form, which may make use of the sentential context. System predictions are judged correct if they exactly match the original derived form.

5.2.2.1 Baseline

As a baseline we considered a trigram model with modified Kneser-Ney smoothing, trained on the training dataset. Each sentence in the testing data was augmented with a set of confabulated sentences, where we replaced a target word with other its derivations or a base form. Unlike the general task, where we generate word forms as character sequences, here we use a set of known inflected forms for each lemma (from the training data). We then use the language model to score the collections of test sentences, and selected the variant with the highest language model score, and evaluate accuracy of selecting the original word form.

5.2.2.2 Encoder–Decoder Model

Here we incorporate the encoder–decoder model that has been introduced in Section 4.3. We replace lemma representations with base form ones and aim to predict their particular context-relevant derived forms.

5.2.2.3 Settings

We used a 3-layer bidirectional LSTM network, with hidden dimensionality $\mathbf{h}$ for both context and lemma form states of 100, and character embedding $\mathbf{c}_j$ of 100.[5] We used pre-trained 300-dimensional Google News word embeddings (Mikolov et al., 2013a,b). During the training of the model, we keep the word embeddings fixed, for greater applicability to unseen test instances. All tokens that didn't appear in this set were replaced with `UNK` sentinel tokens. The network was trained using SGD with momentum until convergence.

5.2.2.4 Results and Error Analysis

We experimented with the encoder–decoder as described in Section 4.3 ("`biLSTM+CTX+BS`"), as well as several variations, namely: excluding context information ("`biLSTM+BS`"), and excluding the bidirectional stem ("`biLSTM+CTX`"). We also investigated how much improvement we can get from knowing the POS tag of the derived form, by presenting it explicitly to the model as extra conditioning context ("`biLSTM+CTX+BS+POS`"). The main motivation for this relates to gerunds, where without the POS, the model often overgenerates nominalisations. We then tried a single-directional context representation, by using only the last hidden states, i.e., $\mathbf{h}^{\rightarrow}_{\text{left}}$ and $\mathbf{h}^{\leftarrow}_{\text{right}}$, corresponding to the words to the immediate left and right of the wordform to be predicted ("`LSTM+CTX+BS+POS`").

We ran two experiments: first, a shared lexicon experiment, where every target base form in the test data was present in the training data; and second, using a split lexicon, where it was *unseen* in the training data. The results are presented in Table 5.1, and show that: (1) context has a strong impact on results, particularly in the shared lexicon case; (2) there is

[5] We also experimented with 15 dimensions, but found this model to perform worse.

	Shared	Split
baseline	0.63	—
biLSTM+BS	0.58	0.36
biLSTM+CTX	0.80	0.45
biLSTM+CTX+BS	0.83	0.52
biLSTM+CTX+BS+POS	0.89	0.63
LSTM+CTX+BS+POS	0.90	0.66

Table 5.1 Accuracy for predicted lemmas (bases and derivations) on shared and split lexicons.

strong complementarity between the context and character representations, particularly in the split lexicon case; and (3) POS information is particularly helpful in the split lexicon case. Note that most of the models significantly outperform our baseline under the shared lexicon setting. The baseline model doesn't support the split lexicon setting (as the derivational forms of interest, by definition, don't occur in the training data), so we cannot generate results in this setting.

We carried out error analysis over the produced forms of the LSTM+CTX+BS+POS model. First, the model sometimes struggles to differentiate between nominal suffixes: in some cases it puts an agentive suffix (-er or -or) in contexts where a non-agentive nominalisation (e.g. -ation or -ment) is appropriate. As an illustration of this, Figure 5.2 is a t-SNE projection of the context representations for simulate vs. simulator vs. simulation, showing that the different nominal forms have strong overlap. Secondly, although the model learns whether to copy or produce a new symbol well, some forms are spelled incorrectly. Examples of this are studint, studion or even studyant rather than student as the agentive nominalisation of study. Here, the issue is opaqueness in the etymology, with student being borrowed from the Old French estudiant. For transformations which are native to English, for example, -ate $\mapsto$ -ation, the model is much more accurate. Table 5.2 shows recall values achieved for various suffix types. We do not present precision since it could not be reliably estimated without extensive manual analysis. In the split lexicon setting, the model sometimes misses double consonants at the end of words, producing wraper and winer and is biased towards generating mostly productive suffixes. An example of the last case might be stoption in place of stoppage. We also

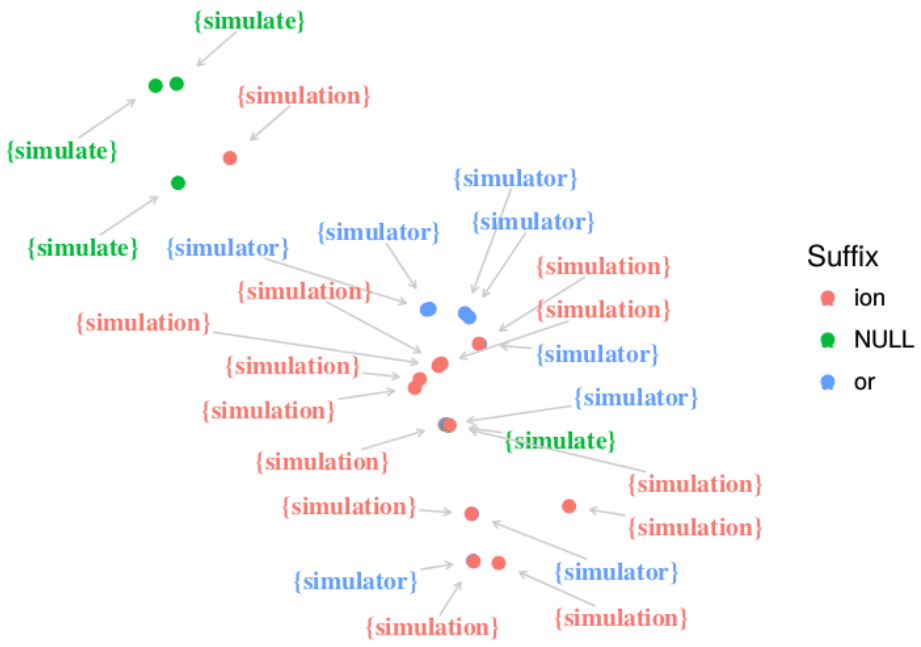


Figure 5.2 An example of t-SNE projection Maaten and Hinton (2008) of context representations for `simulate`.

Affix	$\mathscr{R}$ 1	**Affix**	$\mathscr{R}$ 1	**Affix**	$\mathscr{R}$ 1	**Affix**	$\mathscr{R}$ 1
`-age`	.93	`-al`	.95	`-ance`	.75	`-ant`	.65
`-ation`	.93	`-ator`	.77	`-ee`	.52	`-ence`	.82
`-ent`	.65	`-er`	.87	`-ery`	.84	`-ion`	.93
`-ist`	.80	`-ition`	.89	`-ment`	.90	`-or`	.64
`-th`	.95	`-ure`	.77	`-y`	.83	NULL	.98

Table 5.2 Recall for various suffix types. Here "NULL" corresponds to verb–verb cases.

studied how much the training size affects the model's accuracy by reducing the amount of data. Interestingly, we didn't observe a significant reduction in accuracy. Finally, note that under the split lexicon setting, the model is agnostic of existing derivations, sometimes overgenerating possible forms. A nice illustration of that is `trailation`, `trailment` and `trailer` all being produced in the contexts of `trailer`. In other cases, the model might miss some of the derivations, for instance, predicting only `government` in the contexts of `governance` and `government`. We hypothesise that it is either due to very subtle differences in their contexts, or the higher productivity of `-ment`.

Original	Target Lemma			
transcribe	laptify	fape	crimmle	beteive
transcribe	laptify	fape	crimmle	beterve
transcription	laptification	fapery	crimmler	betention
transcription	laptification	fapication	crimmler	beteption
transcription	laptification	fapionment	crimmler	betention
transcription	laptification	fapist	crimmler	betention
transcription	laptification	fapist	crimmler	beteption
transcript	laptification	fapery	crimmler	betention
transcript	laptification	fapist	crimmler	beteption

Table 5.3 An experiment with nonsense "target" base forms generated in sentence contexts of the "original" word transcribe

Finally, we experimented with some nonsense stems, overwriting sentential instances of transcribe to generate context-sensitive derivational forms. Table 5.3 presents the nonsense stems, the correct form of transcribe for a given context, and the predicted derivational form of the nonsense word. Note that the base form is used correctly (top row) for three of the four nonsense words, and that despite the wide variety of output forms, they resemble plausible words in English. By looking at a larger slice of the data, we observed some regularities. For instance, fapery was mainly produced in the contexts of transcript whereas fapication was more related to transcription. Table 5.3 also shows that some of the stems appear to be more productive than others.

5.2.3 Discussion

We investigated the novel task of context-sensitive derivation prediction for English, and proposed an encoder–decoder model to generate nominalisations. Our best model achieved an accuracy of 90% on a shared lexicon, and 66% on a split lexicon. This suggests that there is regularity in derivational processes and, indeed, in many cases the context is indicative. As we mentioned earlier, there are still many open questions which we leave for future work. Further, we plan to scale to other languages and augment our dataset with Wiktionary data, to realise much greater coverage and variety of derivational forms.

5.3 Conclusion

To conclude, we studied and modelled various classes of English derivations. Our results for contextual modelling confirm observations made by Cotterell et al. (2017b) and suggest that some derivations such as agentivisation are more prototypical and productive than others. Indeed, some derivational suffixes behave similarly to inflectional, i.e., they are widely attached and very few irregularities are observed for them. Our results support the hypothesis that inflections and derivations might belong to a single continuum scale of productivity and restrictedness. Moreover, in the experiments with nonce stems we observe different stem productivity. We do not make attempts to study it in the current thesis, leaving it for future work. Comparison of the results on paradigm-based and contextual derivation suggests that sentential contexts might be more indicative of the form, because NOMLEX categories are more syntactically motivated rather than semantically, therefore, some slots are ambiguous. Finally, we conclude that often derivations require extra information such as base frequency, etymology (if it's not neologism formation) and more fine-grained annotation (making distinctions between RESULT and PROCESS, gender forms, etc.). This motivates construction of new datasets and should also be addressed in future work.

Chapter 6

Conclusions and Future Work

In this thesis, we studied and evaluated various NLP models from a morphological perspective. Specifically, we focused on two types of morphology, inflection and derivation. In terms of the scope of the thesis, the following three questions were addressed:

RQ1: What information do models trained based on the distributional semantics hypothesis capture?

RQ2: Do character-level models provide better representations of morphological similarity than word-based? Which neural architecture better expresses morphological information?

RQ3: How well can derived and inflected forms be predicted directly from a sentential context?

In *Chapter 2* we first provided some background on approaches to morphology modelling existing in linguistics and machine learning. In particular, we discussed the two types of morphology we targeted to model, inflectional and derivational, as well as approaches to their paradigmatic treatment which was further addressed in the thesis. We then continued with a discussion of contemporary distributed models and the distributional semantics principle for learning meaning representations. We outlined that there exists a gap in the evaluation of morphological awareness of the models and very few studies have focused on comparison of various character-level architectures.

Chapter 3 addressed the first two research questions and focused on analysis and evaluation of contemporary neural models. In the first part of the chapter we performed comparison of embeddings obtained in a language modelling task. We assessed a range of word-level and character (*n*-gram)-level neural models in terms of their ability at capturing lexical semantic, morphosyntactic, and morphosemantic binary relations (each relation was represented by means of word vector differences). Evaluating on relations in English and Russian we showed that: (1) models achieved high accuracy in a CLOSED-WORLD setting where each pair represents a relation from our set; while performance significantly dropped in an OPEN-WORLD setting when we augmented the dataset with noisy pairs; (2) character-level models are able to achieve results on par with word-level when provided with less data; (3) morphosyntatic relations are captured much better than morphosemantic and semantic relations; and the gap in performance is even greater in Russian which can be attributed to more transparent word form–meaning relations in this language.

The second part of the chapter focused on machine translation. We compared the performance of word-level and character-level models (character CNN-Highway and BiLSTMs) in encoding the source language linguistic information in Russian $\rightarrow$ English and Estonian $\rightarrow$ English tasks. Our results demonstrated that character-level models improve representation of rare and OOV words. We also observed that BiLSTMs are more focused on word endings, and therefore, they provide more useful information on morphosyntactic similarities (for languages with non-templatic morphology), while CNNs, on the other hand, are better at capturing lemmata. To summarise, we showed that character-level models are superior for low-frequent words. Our results also demonstrate that inflections are more regular in form and meaning than derivations, and character-level models provide a strong signal on types of inflection.

In *Chapter 4* we focused on inflections and partially addressed the third research question, i.e. we studied contextual predictability of inflectional categories. We first formulated a new task of contextual inflection. Unlike traditional morphological inflection, here we aimed at predicting a target word's morphological tags and form from its sentential context and lemma that can be also seen as a special case of a language modelling task. We ran experiments on

18 languages and evaluated several types of models: those that performed direct prediction of the word form, and models that first predicted the target word's tags and then form. We also proposed two settings, inflection of the whole sentence (when only lemmata are provided) and inferring a single word from inflected context. Our results showed that the task is more challenging than morphological inflection, especially for languages with rich morphology. Second, we also achieved lower accuracy in direct form prediction, and models supplied with morphological tags of contextual words performed better. From the perspective of linguistics, some morphological categories such as verbal tense are inherent and are less likely to be predicted from context, while others such as adjective number, case and gender can be inferred from agreement. Finally, we observed that grammatical cases that appear more frequently and at more fixed positions in the sentence are predicted better.[1] This work and results inspired the SIGMORPHON shared task organizing team to run a related sub-task in 2018. The conclusions there were similar, i.e. the task is challenging, and systems perform better when they are presented with morphological tags, although increasing the amount of data, joint training on multiple languages, and more advanced architectures might help to improve the results.

Finally, we continued addressing the third research question in *Chapter 5* focusing on derivations. We additionally discussed paradigmaticity of derivations. We started by describing paradigmatic treatment of derivations (similarly to inflectional paradigms) and continued with contextual prediction. Our study showed that: (1) derivations can be treated paradigmatically; (2) often they present more ambiguous form–meaning mappings and less productivity; (3) there are derivations that are as regular and productive as inflections; (4) results on contextual prediction seem to be better than paradigm-level ones; and (5) finally, we can map both inflections and derivations on a single continuous scale of productivity and specificity. Unlike inflections, much less data exists for derivations, especially cross-lingual, and our results motivate construction of more fine-grained datasets in order to achieve higher accuracies.

[1] Such as the nominative case marking a subject often occurs at the beginning of the sentence in SVO/SOV languages.

6.1 Future work

This part outlines other possible directions of research: diachronic language modelling, usage of morphology models to improve accuracy in low-resource languages, and incorporation of morphology models into the decoder in MT tasks.

6.1.1 Joint Modelling of Etymology and Derivation for English

In terms of the thesis, we only focused on the synchronic view of language but often, especially in derivational modelling, many ambiguous, non-compositional cases can be addressed if we look at language diachronically.

The English lexicon has both Germanic and Latinate origins. For instance, consider the words `student` and `learner`. At their core, both words signify "an agent who seeks knowledge", but their agentive nature is expressed by means of two different suffixes. Can we create a model that will be able to explain the history of derivational morphology as well as predict which of the forms is more likely to appear in a particular time period? To address this question, we need to look back in time and recover the process of their formation.

According to the Merriam-Webster dictionary, the word `learn` appeared before the 12th century (Old English). In contrast, the first entry for `student` dates to the 15th century (Middle English), much after the Norman conquest of England, an important event in the history of the English language. As a direct result of the Norman invasion, a substantial portion of the Middle English lexicon was enriched with words of Latinate stock.

Importantly, the forms might change very significantly over time. For instance, (French) `chaud` "hot" ← (Latin) `calidus` (`[kálidum]` → `[káldum]` → `[kald]` → `[čald]` → `[čaud]` → `[šaud]` → `[šod]` → `[šo]` (`chaud`)), or (English) `lord` ← (Old English) `hlāfweard` "bread keeper". The main principle of historical linguistics postulates that there are no arbitrary and single-word changes, i.e. the change (mainly phonetic) affects all the words of this particular period of time. It is, indeed, a process of phonetic transitions that covers all the words having a corresponding phoneme or combination of phonemes.

We propose to jointly model the processes of diachronic word changes and formation (by means of borrowing and derivation). The goal is to reconstruct the phylogenetic trees and find common patterns and regularities (paradigms) in the transformational processes. More specifically, the probabilistic model should jointly learn possible explanations on how ancient and modern language words are related and evaluate the entire posterior distribution over them.

Such a multilingual diachronic model should find its place in machine translation and other related tasks for resource-poor languages in which current models work less well due to a lack of data. Additional information about related languages and correct identification of cognates should lead to boosts in overall performance.

6.1.2 Low-Resource Language Modelling

Recent technological progress has yielded a significant improvement in machine performance quality in various fields of artificial intelligence such as speech and image recognition, machine translation, dialogue systems, and many others. Part of this success could be attributed to the increase of the power of modern computers and better data availability. In the case of tasks that involve natural language, the success mainly comes from a comparatively small group of well-documented languages such as English, French, German, Russian, Spanish, Italian, and Chinese. A large fraction of less documented languages, although they are widely spoken, is often left out of the scope of current studies. An analysis of re-inflection in 52 languages performed in the context of the SIGMORPHON 2017 shared task (Cotterell et al., 2017a) showed that the accuracy of neural systems drops in low-resource settings (when the number of forms is as low as 100–1000 samples). Synthesising training samples, adding priors to the models as well as the use of data from related languages are possible future directions to address this issue.

Typically, for many languages there is a sufficient amount of unannotated monolingual data, therefore the models are often trained in an unsupervised manner. Availability of annotated data in related languages allows learning the mapping between them. For instance, joint training of high- and low-resource related languages allows knowledge transfer and can

significantly improve the results for the latter in POS tagging (Cotterell and Heigold, 2017). We propose to further investigate this direction with a focus on morphology (and grammar) learning.

Note that many languages do not have their own writing systems, and Duong et al. (2016), Adams et al. (2016a), and Adams et al. (2016b) presented systems of automatic transcription and direct translation of an audio signal, which can certainly be further enriched by models of morphology. Elsewhere Johnson (2008); Johnson et al. (2010) showed that joint modelling of word segmentation and word–object mapping improves the overall performance. The authors also attempted to enrich the model with verbal morphological segmentation but it did not lead to a significant improvement in English. These results could be attributed to the fact that English is a morphologically poor language. The situation should be different in morphologically rich languages where a system has to learn many agreement cases. For instance, in Russian, adjectives should agree with their head nouns. Generally, the system has to segment the utterance, assign the words to adjective and noun classes, infer in which grammatical categories the classes must agree, and how a particular combination of these categories' values is realised in each case. It also has to learn a relative order in which morphemes should be stacked, typically depending on their relevance to the stem morpheme (the more relevant are the closer). We believe enriching the model with morpheme boundaries prediction should improve overall performance. Of course, a successful system would require a generalisation ability and an inference mechanism to be strong enough to capture abstract categories such as grammatical case.

Finally, phonotactic models of extremely low-resource languages (Shcherbakov et al., 2016) that are currently used by field linguists to predict possible lexicon entries will certainly benefit from having a morphological component.

6.1.3 Morphological models for Machine Translation

Another avenue for extension is incorporation of morphology models into the neural decoder of a neural MT system. Generally, we can describe the decoder decision space as translation, transliteration, and copying. And in the case when the sizes of the lexicon and word form set

present in the training data are very limited, word-level translation becomes hard for most languages (especially morphologically rich and resource-poor ones, with polysynthetic as an extreme case of that). Basically, the model has to produce many unseen words itself. But at the very first step the model should choose between utilisation of an existing form and generation of a new one (in many cases by means of deriving it from some other existing form). The decision block initially receives a representation of the meaning to be translated, which is essentially a point in multidimensional space. Many factors, such as sparsity of the corresponding point's neighborhood, could influence its translation decision. Currently these factors are largely unstudied. Our studies on derivational form prediction showed that this is, indeed, very challenging to guess the number of correct possible forms and not over-generate them. Furthermore, in the case of inflection, the produced forms should agree with each other in terms of grammatical case, gender, number, etc. Therefore, the decoder also has to keep track of syntactically connected units (capture hierarchical dependencies) and spread morphological information among them. Having a good model that both captures compositionality and sentence- and word-level hierarchical structures is crucial for many natural language processing tasks if we want to improve open-ended inference and get better generalisation.

Bibliography

Oliver Adams, Graham Neubig, Trevor Cohn, and Steven Bird. Learning a translation model from word lattices. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, pages 2518–2522, 2016a.

Oliver Adams, Graham Neubig, Trevor Cohn, Steven Bird, Quoc Truong Do, and Satoshi Nakamura. Learning a lexicon and translation model from phoneme lattices. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2377–2382, 2016b.

Roee Aharoni, Yoav Goldberg, and Yonatan Belinkov. Improving sequence to sequence learning for morphological inflection generation: The BIU-MIT systems for the SIGMORPHON 2016 shared task for morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 41–48, 2016.

Adam C. Albright. *The identification of bases in morphological paradigms*. PhD thesis, University of California, Los Angeles, 2002.

Inaki Alegria and Izaskun Etxeberria. EHU at the SIGMORPHON 2016 shared task. A simple proposal: Grapheme-to-phoneme for inflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 27–30, 2016.

Inaki Alegria, Izaskun Etxeberria, Mans Hulden, and Montserrat Maritxalar. Porting Basque morphological grammars to Foma, an open-source tool. In *International Workshop on Finite-State Methods and Natural Language Processing*, pages 105–113, 2009.

Stephen R. Anderson. Where's morphology? *Linguistic Inquiry*, 13(4):571–612, 1982.

Stephen R. Anderson. *A-morphous morphology*. Cambridge University Press, 1992.

Alexandra Antonova and Alexey Misyurev. Building a web-based parallel corpus and filtering out machine-translated text. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 136–144, 2011.

Evan L. Antworth. PC-KIMMO: a two-level processor for morphological analysis. *Summer Institute of Linguistics*, 1991.

Mark Aronoff. Word formation in generative grammar. *Linguistic Inquiry Monographs Cambridge, Mass.*, (1):1–134, 1976.

Mark Aronoff and Kirsten Fudeman. *What is morphology?* John Wiley & Sons, 2011.

Mark Aronoff and Mark Lindsay. Productivity, blocking and lexicalization. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Derivational Morphology*, pages 67–83. 2014.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. arXiv:1502.03520 [cs.LG], 2015.

Thierry Artieres et al. Neural conditional random fields. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 177–184, 2010.

Harald R. Baayen, Richard Piepenbrock, and Leon Gulikers. *The CELEX Lexical Data Base on CD-ROM*. Philadelphia: Linguistic Data Consortium, 1993.

R Harald Baayen. 5: Storage and computation in the mental lexicon. In *The mental lexicon*, pages 81–104. BRILL, 2007.

Emmon Bach. *Informal lectures on formal semantics*. Suny Press, 1989.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

L. Bahl and Frederick Jelinek. Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition. *IEEE Transactions on Information Theory*, 21(4):404–411, 1975.

Mark Baker. The mirror principle and morphosyntactic explanation. *Linguistic inquiry*, 16 (3):373–415, 1985.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open information extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 2670–2676, 2007.

Marco Baroni and Alessandro Lenci. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10, 2011.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 23–32, 2012.

Elizabeth Bates and Jeffrey L. Elman. Connectionism and the study of change. *Brain development and cognition: A reader*, pages 623–642, 1993.

Laurie Bauer. *English word-formation*. Cambridge University Press, 1983.

Laurie Bauer. A descriptive gap in morphology. In Geert E. Booij and Jaap Van Marle, editors, *Yearbook of morphology*, volume 1, pages 17–27. 1988.

Laurie Bauer. *Morphological Productivity*, volume 95. Cambridge University Press, 2001.

Robert Beard. *Lexeme-morpheme base morphology: a general theory of inflection and word formation*. Suny Press, 1995.

Robert Beard. Derivation. In Andrew Spencer and Arnold M. Zwicky, editors, *The Handbook of Morphology*, pages 44–65. Blackwell Oxford, 2017.

Kenneth R. Beesley and Lauri Karttunen. Finite-state morphology: Xerox tools and techniques. *CSLI Publications*, 2003.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, 2017.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. Universal dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, 2016.

Robert I. Binnick. *The Oxford handbook of tense and aspect*. Oxford University Press, 2012.

Maximilian Bisani and Hermann Ney. Investigations on joint-multigram models for grapheme-to-phoneme conversion. In *Seventh International Conference on Spoken Language Processing*, 2002.

Barry Blake. *Theories of case*. Cambridge University Press, 2001.

Leonard Bloomfield. *Language*. University of Chicago Press, 1933. Reprint edition (October 15, 1984).

Franz Boas, Helene Boas Yampolsky, and Zellig S. Harris. Kwakiutl grammar with a glossary of the suffixes. *Transactions of the American Philosophical Society*, 37(3):203–377, 1947.

Harry Bochner. *Simplicity in generative morphology*, volume 37. Walter de Gruyter, 2011.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Geert Booij. Inherent versus contextual inflection and the split morphology hypothesis. In Geert E. Booij and Jaap Van Marle, editors, *Yearbook of Morphology*, pages 1–16. Springer, 1996.

Geert Booij. Inflection and derivation. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, volume 5, pages 654–661. Elsevier Science, 2006.

Geert Booij. Paradigmatic morphology. *La Raison Morphologique. Hommage à la Mémoire de Danielle Corbin*, pages 29–38, 2008.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 2787–2795. Curran Associates, Inc., 2013.

Jan Botha and Phil Blunsom. Compositional morphology for word representations and language modelling. In *Proceedings of the International Conference on Machine Learning*, pages 1899–1907, 2014.

Thorsten Brants. TnT: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 224–231, 2000.

John A. Bullinaria and Joseph P. Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526, 2007.

Svetlana Burlak. *Proiskhojdenije jazyka. Fakty, issledovanija, gipotezy*. Litres, 2017.

Miriam Butt. *Theories of case*. Cambridge University Press, 2006.

Joan L Bybee. *Morphology: A study of the relation between meaning and form*, volume 9. John Benjamins Publishing, 1985.

Aoife Cahill. Parsing learner text: to shoehorn or not to shoehorn. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 144–147, 2015.

Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. A gloss composition and context clustering based distributed word sense representation model. *Entropy*, 17(9):6007–6024, 2015a.

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huanbo Luan. Joint learning of character and word embeddings. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015b.

Jason Chiu and Eric Nichols. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370, 2016.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014.

Martin Chodorow and Claudia Leacock. An unsupervised method for detecting grammatical errors. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 140–147, 2000.

Noam Chomsky. *Reflections on language*. 1975.

Noam Chomsky. *Aspects of the Theory of Syntax*, volume 11. MIT Press, 2014.

Noam Chomsky and Morris Halle. *The sound pattern of English*. New York: Harper & Row, 1968.

Grzegorz Chrupała, Georgiana Dinu, and Josef Van Genabith. Learning morphology with Morfette. In *Proceedings of the International Conference on Language Resources and Evaluation*, 2008.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Proceedings of the NIPS 2014 Workshop on Deep Learning*, 2014.

Paul M. Churchland. *The engine of reason, the seat of the soul: A philosophical journey into the brain*. MIT Press, 1996.

Harald Clahsen. Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and brain sciences*, 22(6):991–1013, 1999.

Harald Clahsen and Mayella Almazan. Syntax and morphology in Williams syndrome. *Cognition*, 68(3):167–198, 1998.

Ann Clifton and Anoop Sarkar. Combining morpheme-based machine translation with post-processing morpheme prediction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Volume 1)*, pages 32–42, 2011.

Yael Cohen-Sygal and Shuly Wintner. Finite-state registered automata for non-concatenative morphology. *Computational Linguistics*, 32(1):49–82, 2006.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167, 2008.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

Bernard Comrie. *Tense*. Cambridge University Press, 1985.

Greville Corbett. *Gender*. Cambridge University Press, 1991.

Greville Corbett. *Number*. Cambridge University Press, 2000.

Danielle Corbin. *Morphologie Dérivationnelle et Structuration du Lexique*, volume 193. Walter de Gruyter, 1987.

Marta R. Costa-Jussà and José A.R. Fonollosa. Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 357–361, 2016.

Ryan Cotterell and Georg Heigold. Cross-lingual, Character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, 2017.

Ryan Cotterell and Hinrich Schütze. Joint semantic synthesis and morphological analysis of the derived word. *Transactions of the Association for Computational Linguistics*, 6:33–48, 2018.

Ryan Cotterell, Nanyun Peng, and Jason Eisner. Stochastic contextual edit distance and probabilistic FSTs. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 625–630, 2014.

Ryan Cotterell, Thomas Müller, Alexander Fraser, and Hinrich Schütze. Labeled morphological segmentation with semi-markov models. In *Proceedings of the 19th Conference on Computational Natural Language Learning*, pages 164–174, 2015.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, 2016a.

Ryan Cotterell, Tim Vieira, and Hinrich Schütze. A joint model of orthography and morphological segmentation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 664–669, 2016b.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. The CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, 2017a.

Ryan Cotterell, Ekaterina Vylomova, Huda Khayrallah, Christo Kirov, and David Yarowsky. Paradigm completion for derivational morphology. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 714–720, 2017b.

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

Denis Creissels. Spatial cases. In Andrej Malchukov and Andrew Spencer, editors, *The Oxford handbook of case*, pages 609–625. 2009.

Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1):3, 2007.

James Richard Curran. *From distributional to semantic similarity*. PhD thesis, University of Edinburgh, 2004.

Daniel Dahlmeier and Hwee Tou Ng. A beam-search decoder for grammatical error correction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 568–578, 2012.

Robert Dale and Adam Kilgarriff. Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 242–249, 2011.

Richard J. Daum. Revision of two computer programs for probit analysis. *Bulletin of the ESA*, 16(1):10–15, 1970.

Simon De Deyne and Gert Storms. Word associations: Network and semantic properties. *Behavior Research Methods*, 40(1):213–231, 2008.

Simon De Deyne, Amy Perfors, and Daniel J. Navarro. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1861–1870, 2016.

Sabine Deligne and Frederic Bimbot. Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 169–172, 1995.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

Wolfgang U. Dressler. Prototypical differences between inflection and derivation. *STUF-Language Typology and Universals*, 42(1):3–10, 1989.

Markus Dreyer and Jason Eisner. Graphical models over multiple strings. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (Volume 1)*, pages 101–110, 2009.

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

Markus Dreyer, Jason Smith, and Jason Eisner. Latent-variable modeling of string transductions with finite-state methods. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1080–1089, 2008.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959, 2016.

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

Thomas Eckes and Rüdiger Grotjahn. A closer look at the construct validity of C-tests. *Language Testing*, 23(3):290–325, 2006.

Jason Eisner. Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 1–8, 2002.

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

Jeffrey L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

Jeffrey L. Elman, Elizabeth A. Bates, and Mark H. Johnson. *Rethinking innateness: A connectionist perspective on development*, volume 10. MIT Press, 1998.

Susan M. Ervin. Imitation and structural change in children's language. *New directions in the study of language*, 177(9), 1964.

Nicholas Evans. *A grammar of Kayardild: With historical-comparative notes on Tangkic*, volume 15. Walter de Gruyter, 1995.

Roger Evans and Gerald Gazdar. DATR: a language for lexical knowledge representation. *Computational linguistics*, 22(2):167–216, 1996.

Daniel L. Everett. Pirahã culture and grammar: a response to some criticisms. *Language*, 85 (2):405–442, 2009.

Daniel L. Everett, Brent Berlin, Marco Antonio Gonalves, Paul Kay, Stephen C. Levinson, Andrew Pawley, Alexandre Surralls, Michael Tomasello, and Anna Wierzbicka. Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current anthropology*, 46(4):621–646, 2005.

Manaal Faruqui, Jesse Dodge, Sujay Jauhar, Chris Dyer, Eduard Hovy, and Noah Smith. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 1351–1356, 2015.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. Morphological inflection generation using character sequence to sequence learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 634–643, 2016.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, USA, 1998.

John R. Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

Jerry Fodor and Zenon Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

Jerry Fodor, Thomas Bever, and Garrett Merrill. *The psychology of language: An introduction to psycholinguistics and generative grammar*. McGraw-Hill, 1974.

Alan Ford, Rajendra Singh, and Gita Martohardjono. *Pace Pānini: Towards a word-based theory of morphology*, volume 34. Peter Lang Pub Incorporated, 1997.

Michael Fortescue, Marianne Mithun, and Nicholas Evans. *The Oxford handbook of polysynthesis*. Oxford University Press, 2017.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. Modeling inflection and word-formation in SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 664–674, 2012.

Daniel Fried and Kevin Duh. Incorporating both distributional and relational semantics in word representations. In *Proceedings of the International Conference on Learning Representations (workshop contribution)*, 2015.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1199–1209, 2014.

Yasuhisa Fujii, Kazumasa Yamamoto, and Seiichi Nakagawa. Deep-hidden conditional neural fields for continuous phoneme speech recognition. In *Proceedings International Workshop of Statistical Machine Learning for Speech Processing*, 2012.

Philip Gage. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38, 1994.

Maayan Geffet and Ido Dagan. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 107–114, 2005.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. SemEval-2007 Task 4: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluation (SemEval 2007)*, pages 13–18, 2007.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 8–15, 2016.

Gene H. Golub and Christian Reinsch. Singular value decomposition and least squares solutions. *Numerische mathematik*, 14(5):403–420, 1970.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation*, 2018.

Alex Graves, Santiago Fernández, Jürgen Schmidhuber, Janusz Kacprzyk, Erkki Oja, and Sławomir Zadrożny. Bidirectional LSTM networks for improved phoneme classification and recognition. In *Artificial Neural Networks: Formal Models and Their Applications*, pages 799–804, 2005.

Joseph Harold Greenberg. *Universals of language*. London: MIT Press, 1963.

Jacob Grimm. *Deutsche grammatik*, volume 3. C. Bertelsmann, 1890.

Jane Grimshaw. *Argument structure*. MIT Press, 1990.

Emiliano Guevara. Computing semantic compositionality in distributional semantics. In *Proceedings of the 9th International Conference on Computational Semantics*, pages 135–144, 2011.

Jan Hajic, Martin Cmejrek, Bonnie Dorr, Yuan Ding, Jason Eisner, Daniel Gildea, Terry Koo, Kristen Parton, Gerald Penn, Dragomir Radev, and Owen Rambow. Natural language generation in the context of machine translation. In *Summer workshop final report, Johns Hopkins University*, 2002.

Huda Hakami, Kohei Hayashi, and Danushka Bollegala. Why does pairdiff work? A mathematical analysis of bilinear relational compositional operators for analogy detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2493–2504, 2018.

Morris Halle and Alec Marantz. Some key features of distributed morphology. *MIT working papers in linguistics*, 21(275):88, 1994.

Zellig S. Harris. Morpheme alternants in linguistic analysis. *Language*, 18:169–180, 1942.

Zellig S. Harris. From morpheme to utterance. In *Papers on Syntax*, pages 45–70. Springer, 1946.

Zellig S. Harris. Distributional structure. In *Papers on Syntax*, pages 3–22. Springer, 1981.

Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie. *The world atlas of language structures*. Oxford University Press, 2005.

Marc Hauser, Noam Chomsky, and Tecumseh Fitch. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579, 2002.

Jennifer Hay and Harald Baayen. Parsing and productivity. In *Yearbook of morphology 2001*, pages 203–235. Springer, 2002.

Jennifer Hay and Ingo Plag. What constrains possible suffix combinations? on the interaction of grammatical and processing restrictions in derivational morphology. *Natural Language & Linguistic Theory*, 22(3):565–596, 2004.

Georg Heigold, Guenter Neumann, and Josef van Genabith. An extensive empirical evaluation of character-based morphological tagging for 14 languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 505–513, 2017.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 Task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, pages 33–38, 2010.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Burry Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

Cong Duy Vu Hoang, Trevor Cohn, and Gholamreza Haffari. Incorporating side information into recurrent neural network language models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1250–1255, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term memory. *Neural computation*, 9(8):1735–1780, 1997.

Charles F. Hockett. Problems of morphemic analysis. *Language*, 23(4):321–343, 1947.

Charles F. Hockett. Two models of grammatical description. *Word*, 10(2-3):210–234, 1954.

Charles F. Hockett. *A Course In Modern Linguistics*. The MacMillan Company, 1958.

Charles F. Hockett. *The view from language: Selected essays, 1948-1974*. University of Georgia Press, 1977.

Charles F. Hockett and Stuart A. Altmann. A note on design features. *Animal communication: Techniques of study and results of research*, pages 61–72, 1968.

Charles F. Hockett and Charles D. Hockett. The origin of speech. *Scientific American*, 203 (3):88–97, 1960.

Matthew Honnibal and Mark Johnson. Joint incremental disfluency detection and dependency parsing. *Transactions of the Association for Computational Linguistics*, 2:131–142, 2014.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: (Volume 1 : Long Papers)*, pages 873–882, 2012.

Matthias Huck, Aleš Tamchyna, Ondřej Bojar, and Alexander Fraser. Producing unseen morphological variants in statistical machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 369–375, 2017.

Mans Hulden, Markus Forsberg, and Malin Ahlberg. Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 569–578, 2014.

Larry M. Hyman. Suffix ordering in Bantu: A morphocentric approach. In *Yearbook of morphology*, pages 245–281. Springer, 2003.

Ray Jackendoff. Morphological and semantic regularities in the lexicon. *Language*, pages 639–671, 1975.

Jeri J Jaeger, Alan H Lockwood, David L Kemmerer, Robert D Van Valin Jr, Brian W Murphy, and Hanif G Khalak. A positron emission tomographic study of regular and irregular verb morphology in english. *Language*, pages 451–497, 1996.

Roman Jakobson. *Zur struktur der russischen verbums*. Pražskỳ Linguistickỳ Kroužek, 1932.

Roman Jakobson. *Russian and Slavic grammar: studies 1931-1981*, volume 106. Walter de Gruyter, 2011.

Fred Jelinek, Robert L. Mercer, Lalit R. Bahl, and James K. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 1977.

Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 372–379, 2007.

Mark Johnson. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 398–406, 2008.

Mark Johnson, Katherine Demuth, Bevan Jones, and Michael J. Black. Synergies in learning words and their referents. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, pages 1018–1026. Curran Associates, Inc., 2010.

Michael I. Jordan. Serial order: A parallel distributed processing approach. *Advances in psychology*, 121:471–495, 1997.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. SemEval-2012 Task 2: Measuring degrees of relational similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364, 2012.

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, 2013.

Katharina Kann and Hinrich Schütze. Single-model encoder-decoder with explicit morphological representation for reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 555–560, 2016.

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

Ronald M Kaplan and Martin Kay. Regular models of phonological rule systems. *Computational linguistics*, 20(3):331–378, 1994.

Lauri Karttunen and Kenneth R Beesley. *Two-level rule compiler*. Xerox Corporation. Palo Alto Research Center, 1992.

Lauri Karttunen and Kenneth R Beesley. A short history of two-level morphology. *ESSLLI-2001 Special Event titled "Twenty Years of Finite-State Morphology"*, 2001.

Yova Kementchedjhieva, Johannes Bjerva, and Isabelle Augenstein. Copenhagen at CoNLL–SIGMORPHON 2018: Multilingual inflection in context with explicit morphosyntactic decoding. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 93–98, 2018.

Aleksandr E. Kibrik. Archi (Caucasian – Daghestanian). In Andrew Spencer and Arnold M. Zwicky, editors, *The Handbook of Morphology*, pages 455–476. Blackwell Oxford, 1998.

Joo-Kyung Kim and Marie-Catherine de Marneffe. Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1625–1630, 2013.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander Rush. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

David King. Evaluating sequence alignment for learning inflectional morphology. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 49–53, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

George Anton Kiraz. Multitiered nonlinear morphology using multitape finite automata: a case study on syriac and arabic. *Computational Linguistics*, 26(1):77–105, 2000.

Maxim Kireev, Natalia Slioussar, Alexander D. Korotkov, Tatiana V. Chernigovskaya, and Svyatoslav V. Medvedev. Changes in functional connectivity within the fronto-temporal brain network induced by regular and irregular russian verb production. *Frontiers in Human Neuroscience*, 9, 2015.

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. Very-large scale parsing and normalization of wiktionary morphological paradigms. 2016.

Max Kisselew, Sebastian Padó, Alexis Palmer, and Jan Šnajder. Obtaining a better understanding of distributional models of German derivational morphology. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 58–63, 2015.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of the Association for Computational Linguistics, System Demonstrations*, pages 67–72, 2017.

Christine Klein-Braley and Ulrich Raatz. A survey of research on the C-test. *Language Testing*, 1(2):134–146, December 1984.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, and Chris Dyer. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, 2007.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Maximilian Köper, Christian Scheible, and Sabine Schulte im Walde. Multilingual reliability and "semantic" structure of continuous word spaces. In *Proceedings of the Eleventh International Workshop on Computational Semantics (IWCS-11)*, pages 40–45, 2015.

Livia Körtvélyessy. Evaluative derivation. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Derivational Morphology*. 2014.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16:359–389, 2010.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 1097–1105. Curran Associates, Inc., 2012.

Jerzy Kuryłowicz. *Dérivation lexicale et dérivation syntaxique (Contribution à la théorie des parties du discours)*. 1936.

Jerzy Kuryłowicz. *The inflectional categories of Indo-European*, volume 3. Adler's Foreign Books Inc, 1964.

Marta Kutas and Steven A. Hillyard. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, 1980.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*, 2001.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the International Conference on Machine Learning*, pages 2879–2888, 2018.

Robin Lakoff. What you can do with words: Politeness, pragmatics and performatives. In *Proceedings of the Texas Conference on Performatives, Presuppositions and Implicatures*, pages 79–106, 1977.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, 2016.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. Compositionally derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1517–1526, 2013.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. Automated grammatical error detection for language learners. *Synthesis lectures on human language technologies*, 7(1):1–170, 2014.

Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Alessandro Lenci and Giulia Benotto. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, pages 75–79, 2012.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the 18th conference on Computational Natural Language Learning*, pages 171–180, 2014a.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 2177–2185. Curran Associates, Inc., 2014b.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015a.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, 2015b.

Rochelle Lieber. *On the organization of the lexicon*. PhD thesis, Massachusetts Institute of Technology, 1980.

Rochelle Lieber. *Deconstructing morphology: Word formation in syntactic theory*. University of Chicago Press, 1992.

Rochelle Lieber. *Morphology and Lexical Semantics*, volume 104. Cambridge University Press, 2004.

Rochelle Lieber and Pavol Štekauer. *The Oxford handbook of derivational morphology*. Oxford Handbooks in Linguistics, 2014.

Wang Ling, Chris Dyer, Alan W. Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530, 2015a.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan Black. Character-based neural machine translation. *arXiv preprint arXiv:1511:04586*, 2015b.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

Ling Liu and Lingshuang Jack Mao. Morphological reinflection with conditional random fields and unsupervised features. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 36–40, 2016.

Thang Luong, Richard Socher, and Christopher Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, 2013.

Christopher Lyons. *Definiteness*. Cambridge University Press, 1999.

Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1064–1074, 2016.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.

David Maier. The complexity of some problems on subsequences and supersequences. *Journal of the ACM (JACM)*, 25(2):322–336, 1978.

Peter Makarov and Simon Clematide. Neural transition-based string transduction for limited-resource setting in morphology. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 83–93, 2018.

Márton Makrai, Dávid Nemeskey, and András Kornai. Applicative structure in vector space models. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 59–63, 2013.

Andrej Malchukov and Andrew Spencer. *The Oxford handbook of case*. Oxford University Press, 2009.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. SemEval-2010 Task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, 2010.

Gary F. Marcus. Can connectionism save constructivism? *Cognition*, 66(2):153–182, 1998.

Gary F. Marcus. *The algebraic mind: Integrating connectionism and cognitive science*. MIT Press, 2003.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2): 313–330, 1993.

Jaap van Marie. Paradigms. In *Encyclopedia of Language and Linguistics*, pages 2927–2930. Oxford, 1994.

William D. Marslen-Wilson and Lorraine K. Tyler. Dissociating types of mental computation. *Nature*, 387(6633):592, 1997.

William J. Masek and Michael S. Paterson. A faster algorithm computing string edit distances. *Journal of Computer and System sciences*, 20(1):18–31, 1980.

Peter H. Matthews. Some concepts in word-and-paradigm morphology. *Foundations of Language*, pages 268–289, 1965.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. The NomBank Project: An interim report. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, 2004.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at the International Conference on Learning Representations*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 3111–3119. Curran Associates, Inc., 2013b.

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

Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429, 2010.

Andriy Mnih and Geoffrey Hinton. Three new graphical models for statistical language modelling. In *Proceedings of the 24th International Conference on Machine learning*, pages 641–648. ACM, 2007.

Andriy Mnih and Geoffrey E. Hinton. A scalable hierarchical distributed language model. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, pages 1081–1088. Curran Associates, Inc., 2009.

Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 2265–2273. Curran Associates, Inc., 2013.

Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of the 29th International Conference on Machine Learning*, volume 2, pages 1751–1758, 2012.

Mehryar Mohri. Semiring frameworks and algorithms for shortest-distance problems. *Journal of Automata, Languages and Combinatorics*, 7(3):321–350, 2002.

Mehryar Mohri, Fernando Pereira, and Michael Riley. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88, 2002.

Arlene Moskowitz. On the status of vowel shift in English. In *Cognitive Development and the Acquisition of Language*. Academic Press, 1973.

Thomas Müller and Hinrich Schütze. Robust morphological tagging with word representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 526–536, 2015.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, 2013.

Ryo Nagata and Keisuke Sakaguchi. Phrase structure annotation and parsing for learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1837–1847, 2016.

Silvia Necşulescu, Sara Mendes, David Jurgens, Núria Bel, and Roberto Navigli. Reading between the lines: Overcoming data sparsity for accurate classification of lexical relationships. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 182–192, 2015.

Graham Neubig, Taro Watanabe, Shinsuke Mori, and Tatsuya Kawahara. Substring-based machine translation. *Machine translation*, 27(2):139–166, 2013.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, 2014.

Garrett Nicolai, Colin Cherry, and Grzegorz Kondrak. Inflection generation as discriminative string transduction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 922–931, 2015.

Garrett Nicolai, Bradley Hauer, Adam St. Arnaud, and Grzegorz Kondrak. Morphological reinflection via discriminative string transduction. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 31–35, 2016.

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

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, and Liesbeth et al. Augustinus. Universal dependencies 2.1, 2017. URL `http://hdl.handle.net/11234/1-2515`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (Volume 1)*, pages 160–167, 2003.

Jose Oncina and Marc Sebban. Learning stochastic edit distance: Application in handwritten character recognition. *Pattern recognition*, 39(9):1575–1587, 2006.

Robert Östling. Morphological reinflection with convolutional neural networks. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 23–26, 2016.

Sebastian Padó, Aurélie Herbelot, Max Kisselew, and Jan Šnajder. Predictability of distributional semantics in derivational word formation. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 1285–1297, 2016.

Patrick Pantel, Deepak Ravichandran, and Eduard Hovy. Towards terascale semantic acquisition. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 771–777, 2004.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Barbara H. Partee. Mathematical methods in linguistics. *Recherche*, 67:02, 1990.

Jian Peng, Liefeng Bo, and Jinbo Xu. Conditional neural fields. In *Proceedings of the Neural Information Processing Systems*, pages 1419–1427, 2009.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVE: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

Slav Petrov and Ryan McDonald. Overview of the 2012 shared task on parsing the web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*, volume 59, 2012.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 2012.

Steven Pinker and Ray Jackendoff. The faculty of language: what's special about it? *Cognition*, 95(2):201–236, 2005.

Steven Pinker and Alan Prince. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1):73–193, 1988.

Ingo Plag and Harald Baayen. Suffix ordering and morphological processing. *Language*, pages 109–152, 2009.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, 2016.

Frans Plank. Inflection and derivation. In Stephen Walker and JMY Simpson, editors, *Encyclopedia of Language and Linguistics*, volume 3, pages 1671–1678. Pergamon, 1994.

Tony Plate. Holographic reduced representations: Convolution algebra for compositional distributed representations. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 30–35, 1991.

Kim Plunkett and Virginia Marchman. From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, 48(1):21–69, 1993.

Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

Amanda Pounder. *Process and paradigms in word-formation morphology*, volume 131. Walter de Gruyter, 2011.

Jan Ptáček and Zdeněk Žabokrtský. Synthesis of Czech sentences from tectogrammatical trees. In *International Conference on Text, Speech and Dialogue*, pages 221–228, 2006.

Randolph Quirk and Charles Wrenn. *An old English grammar*. Routledge, 2002.

Mohammad Sadegh Rasooli and Joel R Tetreault. Non-monotonic parsing of fluent umm i mean disfluent sentences. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 48–53, 2014.

Pushpendre Rastogi, Ryan Cotterell, and Jason Eisner. Weighting finite-state transductions with neural context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 623–633, 2016.

K. Rice. *The Athabaskan Verb*. Cambridge, 2000.

John Richardson. Lexical derivation. *Journal of Psycholinguistic Research*, 6(4):319–336, 1977.

Laura Rimell. Distributional lexical entailment by topic coherence. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 511–519, 2014.

Eric Sven Ristad and Peter N Yianilos. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532, 1998.

Stephen Roller and Katrin Erk. Relations such as Hypernymy: Identifying and exploiting Hearst patterns in distributional vectors for lexical entailment. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2163–2172, 2016.

Stephen Roller, Katrin Erk, and Gemma Boleda. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1025–1036, 2014.

Andrew Rosenberg and Julia Hirschberg. V-Measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2007 (EMNLP-CoNLL 2007)*, pages 410–420, 2007.

David E. Rumelhart and James L. McClelland. Parallel distributed processing, explotation in the microstructure of cognition (Volume 1 : Foundations). *Computational Models of Cognition and Perception*, 1987.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. In David E. Rumelhart and James L. Mcclelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Volume 1: Foundations)*, pages 318–362. 1986.

Pauliina Saarinen and Jennifer Hay. Affix ordering in derivation. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford Handbook of Derivational Morphology*. 2014.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15, 2002.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. Error-repair dependency parsing for ungrammatical texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–195, 2017.

Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798, 2007.

Bahar Salehi, Paul Cook, and Timothy Baldwin. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, 2014.

Cicero D. Santos and Bianca Zadrozny. Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1818–1826, 2014.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. Chasing hypernyms in vector spaces with entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 38–42, 2014.

Sunita Sarawagi and William W. Cohen. Semi-markov conditional random fields for information extraction. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, pages 1185–1192. MIT Press, 2005.

Ferdinand de Saussure. *Course in general linguistics (W. Baskin, Trans.)*. New York: Philosophical Library, 1959.

Sergio Scalise. Inflection and derivation. *Linguistics*, 26(4):561–582, 1988.

Hinrich Schütze. Word space. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems*, pages 895–902. Morgan-Kaufmann, 1993.

Marcel Paul Schützenberger. On the definition of a family of automata. *Information and control*, 4(2-3):245–270, 1961.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

Ilya Segalovich. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*, pages 273–280, 2003.

Ann Senghas, Sotaro Kita, and Asli Özyürek. Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science*, 305(5691):1779–1782, 2004.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, 2016.

Andrei Shcherbakov, Ekaterina Vylomova, and Nick Thieberger. Phonotactic modeling of extremely low resource languages. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 84–93, 2016.

Anna Siewierska. *Person*. Cambridge University Press, 2004.

Maria Singson, Diana Mahony, and Virginia Mann. The relation between reading ability and morphological skills: Evidence from derivational suffixes. *Reading and writing*, 12(3): 219–252, 2000.

Friedrich Slotty. Das problem der wortarten. *Forschungen und Fortschritte*, 7, 1932.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216, 1990.

Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Learning continuous phrase representations and syntactic parsing with recursive neural networks. In *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*, pages 1–9, 2010.

Richard Socher, Cliff C. Lin, Chris Manning, and Andrew Y. Ng. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 129–136, 2011.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211, 2012.

Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems 25 (NIPS-13)*, 2013a.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, volume 1631, page 1642, 2013b.

Linfeng Song, Xiaochang Peng, Yue Zhang, Zhiguo Wang, and Daniel Gildea. AMR-to-text generation with synchronous node replacement grammar. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 7–13, 2017.

Aleksei Sorokin, Tatiana Shavrina, Olga Lyashevskaya, Viktor Bocharov, Svetlana Alexeeva, Kira Droganova, Alena Fenogenova, and Dmitry Granovsky. MorphoRuEval-2017: An evaluation track for the automatic morphological analysis methods for Russian. *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, 1:297–313, 2017.

Andrew Spencer. *Morphological Theory: An Introduction to Word Structure in Generative Grammar*. Wiley-Blackwell, 1991.

Richard Sproat and Osamu Fujimura. Allophonic variation in English and its implications for phonetic implementation. *Journal of phonetics*, 21(3):291–311, 1993.

Rupesh Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 2377–2385. Curran Associates, Inc., 2015.

Susan Steele. Towards a theory of morphological information. *Language*, pages 260–309, 1995.

Gregory T. Stump. A paradigm-based theory of morphosemantic mismatches. *Language*, pages 675–725, 1991.

Gregory T. Stump. *Inflectional morphology: A theory of paradigm structure*, volume 93. Cambridge University Press, 2001.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 3104–3112. Curran Associates, Inc., 2014.

John Sylak-Glassman, Christo Kirov, Matt Post, Roger Que, and David Yarowsky. A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 72–93. Springer, 2015a.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, 2015b.

Bogdan Szymanek. *A panorama of Polish word-formation*. Wydawn. KUL, 2010.

Dima Taji, Ramy Eskander, Nizar Habash, and Owen Rambow. The Columbia University-New York University Abu Dhabi SIGMORPHON 2016 morphological reinflection shared task submission. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 71–75, 2016.

Aleš Tamchyna, Marion Weller-Di Marco, and Alexander Fraser. Modeling target-side inflection in neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 32–42, 2017.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison Wesley, 2006a.

Yee Fan Tan, Min-Yen Kan, and Hang Cui. Extending corpus-based identification of light verb constructions using a supervised learning framework. In *Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a Multilingual Context*, pages 49–56, 2006b.

Shane Templeton. Spelling, phonology, and the older student. *Developmental and cognitive aspects of learning to spell: A reflection of word knowledge*, pages 85–96, 1980.

Edward Lee Thorndike. *The teaching of English suffixes*, volume 847. Teachers College, Columbia University, 1941.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the ACL (ACL 2010)*, pages 384–394, 2010.

Peter D. Turney. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416, 2006.

Michael T Ullman. The declarative/procedural model of lexicon and grammar. *Journal of psycholinguistic research*, 30(1):37–69, 2001.

Michael T Ullman. Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, 92(1-2):231–270, 2004.

Lev Uspensky. *A word about word*. Govt. publisher of the USSR, 1956.

Vladimir Uspensky. *Trudy po nematematike*. OGI, 2002.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

Clara Vania and Adam Lopez. From characters to words to in between: Do we capture morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2016–2027, 2017.

David Vilar, Jan-T. Peter, and Hermann Ney. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 33–39, 2007.

Antoine Vinel, Trinh Minh Tri Do, and Thierry Artieres. Joint optimization of hidden conditional random fields and non linear feature extraction. In *Document Analysis and Recognition, Proceedings of the 2011 International Conference on*, pages 513–517, 2011.

Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions Information Theory*, 13(2):260–269, 1967.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, 2007.

Pavol Štekauer. Derivational paradigms. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford handbook of derivational morphology*, chapter 12, pages 354–369. Oxford University Press, Oxford, 2014.

Ekaterina Vylomova, Trevor Cohn, Xuanli He, and Gholamreza Haffari. Word representation models for morphologically rich languages in neural machine translation. *arXiv preprint arXiv:1606.04217*, 2016.

Géraldine Walther and Benoît Sagot. Developing a large-scale lexicon for a less-resourced language: General methodology and preliminary experiments on Sorani Kurdish. In *Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop)*, 2010.

Géraldine Walther, Guillaume Jacques, and Benoît Sagot. Uncovering the inner architecture of Khaling verbal morphology. In *Proceedings of the Third Workshop on Sino-Tibetan languages of Sichuan*, 2013.

Mengqiu Wang and Christopher D. Manning. Effect of non-linear deep architecture in sequence labeling. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 1285–1291, 2013.

Peilu Wang, Yao Qian, Frank K. Soong, Lei He, and Hai Zhao. Part-of-speech tagging with bidirectional long short-term memory recurrent neural network. *arXiv preprint arXiv:1510.06168*, 2015.

Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. Learning to distinguish hypernyms and co-hyponyms. 2014.

Gerhard Weikum and Martin Theobald. From information to knowledge: harvesting entities and relationships from web sources. In *Proceedings of the Twenty Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 65–76, 2010.

Richard Wicentowski and David Yarowsky. *Modeling and learning multilingual inflectional morphology in a minimally supervised framework*. PhD thesis, Johns Hopkins University, 2000.

Anna Wierzbicka. *The Semantics of Grammar*, volume 18. John Benjamins Publishing, 1988.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

Shuangzhi Wu, Dongdong Zhang, Ming Zhou, and Tiejun Zhao. Efficient disfluency detection with transition-based parsing. In *Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 495–503, 2015.

Chang Xu, Yanlong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. RC-NET: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM Conference on Information and Knowledge Management (CIKM 2014)*, pages 1219–1228, 2014.

Ichiro Yamada, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, Masaki Murata, Stijn De Saeger, Francis Bond, and Asuka Sumida. Hypernym discovery based on distributional similarity and hierarchical structures. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 929–937, 2009.

Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. Comparative study of CNN and RNN for natural language processing. *arXiv preprint arXiv:1702.01923*, 2017.

Masashi Yoshikawa, Hiroyuki Shindo, and Yuji Matsumoto. Joint transition-based dependency parsing and disfluency detection for automatic speech recognition texts. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1036–1041, 2016.

Mo Yu and Mark Dredze. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 545–550, 2014.

Andrei Zaliznyak. *Russkoje imennoje slovoizmenenie (Russian nominal inflection)*. Moscow: Nauka, 1967.

Richard Zens and Hermann Ney. Improvements in phrase-based statistical machine translation. In *Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2004.

Alisa Zhila, Wen-tau Yih, Christopher Meek, Geoffrey Zweig, and Tomas Mikolov. Combining heterogeneous models for measuring relational similarity. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1000–1009, 2013.