Word order variation and dependency length minimisation : a cross-linguistic computational approach

Gulordava, Kristina

UNIVERSITÉ
DE GENÈVE
FACULTÉ DES LETTRES

Doctoral Dissertation

Word order variation and dependency length minimisation

A cross-linguistic computational approach

Kristina Gulordava

Supervisor: Prof. Paola Merlo

Abstract

Word order is one of the most readily observed and extensively studied aspects of the human language. The central object of study of this thesis are cases of variation in word order, i.e., cases when one syntactic structure can be expressed using more than one grammatical linearisation. We are interested in cross-linguistical properties of word order variation and, in particular, in phenomena related to dependency length minimisation (DLM). DLM is known as a tendency for words and phrases that are close in the syntactic structure (dependents) to be linearly adjacent. The evidence for this principle was observed in many languages of the world and in various types of word order distributions.

We analyse DLM phenomena observed in word order variation using a computational approach. Our work capitalises on syntactically-annotated corpora (treebanks) and statistical methods which are essential for drawing generalisations from word order variation data across dozens of languages. To analyse similar constructions in various languages in the same way, we treat word order as a mapping between the syntactic structures of utterances, provided by the treebanks, and their linearisations. Since we use treebanks which annotate different languages starting from the same syntactic criteria, the distributions of word order mappings extracted from these treebanks can be compared meaningfully to each other.

This thesis presents three cross-linguistic computational studies of word order variation and dependency length minimisation at three levels of linguistic representation. First, we look at word order and dependency length distributions at the language level. One of the aims of this study is to examine the general formulation of the DLM

principle applied to all types of constructions and dependency relations in a language. All languages tend to minimise dependency lengths; however, the degree of this minimisation varies substantially. The measure of the rate of DLM at the language level provides a way to compare languages typologically across a new interesting dimension.

Secondly, we zoom in on the DLM effects in word order distributions in one syntactic construction: adjective variation in Romance languages. We formalise the predictions of the language-level global DLM principle for this complex syntactic construction involving several dependencies. We test these predictions systematically in treebanks of five Romance languages. We reveal several DLM-related patterns in adjective placement, confirming the promising approach to formalising and probing DLM. For instance, we find that adjectives tend to appear before the noun they modify when the noun has an additional right dependent than when there is no such dependent. We also highlight the limitations of the global DLM principle, e.g., that it cannot explain the fact that different dependencies are optimised to a different extent.

Finally, we analyse distributions of word order as generalisations of linearisation decisions at production time. To this end, we develop a linearisation system which models online, word-by-word production of word order. It is conceived as a plausible model of the word order production process and, at the same time, as a model of word order distributions both at the language and at the construction level. This model integrates the choice between two options for the cases of word order variation and conditions these choices on dependency length factors.

The contributions of this thesis are relevant, first of all, to the linguistic work interested in questions about word order and DLM. Additionally, this thesis is tightly linked to the research in natural language processing (NLP). As part of the analysis of word order variation and DLM at the language level, we investigate how these properties affect the performance of statistical parsers. Our linearisation model is related to the previous work in natural language generation and sentence linearisation and is evaluated against a state-of-the-art NLP system. The results of this thesis are, therefore, of interest to computational studies of syntax and variation and the field of natural language processing.

Resumé

L'ordre des mots est l'une des propriétés les plus facilement observées et les plus étudiées du langage humain. L'objet central de cette thèse est l'étude des cas de variation d'ordre de mots, i.e. des cas où une structure syntaxique donnée peut être exprimée en utilisant plus d'une linéarisation grammaticale. Nous nous intéressons aux propriétés inter-linguistiques de la variation de l'ordre des mots et, en particulier, aux phénomènes de minimisation de la longueur des dépendances (DLM). La DLM est la propriété des mots et des syntagmes proches dans la structure syntaxique d'être adjacents lorsque l'on considère l'ordre linéaire de la phrase. Ce principe est observable dans de nombreuses langues et pour divers types de distributions d'ordres de mots.

Nous analysons les phénomènes de DLM observés dans la variation de l'ordre des mots en utilisant une approche computationnelle. Notre travail utilise des corpus annotés en syntaxe (représentations arborescentes) et des méthodes statistiques. Les deux sont essentiels pour obtenir des généralisations à partir de données annotées pour des dizaines de langues. Pour analyser de la même façon des constructions similaires dans différentes langues, nous traitons l'ordre des mots comme une correspondance entre les structures syntaxiques, représentées sous forme d'arbres, et leurs linéarisations. Puisque nous utilisons des corpus arborés qui annotent différentes langues en utilisant les mêmes critères syntaxiques, les distributions inter-linguistiques de l'ordre des mots extraites de ces arbres peuvent être comparés de manière significative.

Cette thèse présente trois études computationnelles inter-linguistiques de la variation

de l'ordre des mots et de la minimisation de la longueur des dépendances à trois niveaux de représentation linguistique. Premièrement, nous examinons l'ordre des mots et les distributions de la longueur des dépendances au niveau de la langue. L'un des objectifs de cette étude est d'examiner la formulation générale du principe de la DLM appliqué à tous les types de relations syntaxiques d'une langue. Toutes les langues ont tendance à minimiser la longueur des dépendances. Cependant, l'ampleur de cette minimisation varie considérablement. Mesurer l'étendue de la DLM au niveau de la langue permet de comparer les langues à travers une nouvelle dimension typologique intéressante.

Deuxièmement, nous nous concentrons sur les effets de DLM dans les distributions de l'ordre des mots d'une construction syntaxique choisie: la variation du placement des adjectifs dans les langues romanes. Nous formalisons les prédictions du principe de DLM globale pour cette construction syntaxique complexe et impliquant plusieurs dépendances. Nous testons systématiquement ces prédictions dans des corpus arborés de cinq langues romanes. Nous révélons plusieurs patterns dans le placement des adjectifs liés au DLM, ce qui confirme notre approche prometteuse pour formaliser et examiner la DLM. Par exemple, nous constatons que l'adjectif a tendance à apparaître plus fréquemment avant le nom qu'il modifie, quand le nom a une dépendance supplémentaire à droite (par exemple, un syntagme prépositionnel), que lorsqu'il n'y a pas ce genre de dépendances. Nous soulignons également les limites du principe de la DLM globale. Par exemple, il ne peut pas expliquer le fait que différentes dépendances sont minimisées de manière différente.

Enfin, nous analysons les distributions de l'ordre des mots comme des généralisations de décisions de linéarisation au moment de la production. Pour cela, nous développons un système de linéarisation en ligne qui modélise la production de l'ordre des mots, mot par mot. Il est conçu comme un modèle plausible du processus psychologique de production de l'ordre des mots et, en même temps, comme un modèle de distribution de l'ordre des mots au niveau de la langue et au niveau des constructions syntaxiques. Ce modèle intègre le choix entre deux options pour les cas de variation de l'ordre des mots et conditionne ce choix en s'appuyant sur le facteur de longueur de dépendances.

Les contributions de cette thèse sont pertinentes, tout d'abord, d'un point de vue linguistique. De plus, cette thèse est liée à la recherche en traitement automatique du langage naturel (TALN). Dans le cadre de l'analyse de la variation de l'ordre des mots et de la DLM au niveau du langage, nous étudions comment ces propriétés influencent les performances des analyseurs statistiques. Notre modèle de linéarisation est à mettre en relation avec les travaux précédents sur la génération automatique de phrases d'une langue naturelle. Il est évalué par rapport à un système de TALN à l'état de l'art. Les résultats de cette thèse apportent donc des contributions en syntaxe computationnelle et dans le domaine de la variation de l'ordre des mots, ainsi que pour les études liées au traitement automatique du langage naturel.

Acknowledgements

This thesis has benefited greatly from the collaborations, discussions and support of many people whom I would like to thank here.

First, I would like to thank my supervisor, Paola Merlo, for drawing me into the world of computational syntax and language universals. My published work and this thesis, in particular, came to be more rigorous, comprehensible and contentful because of the enthusiasm and countless hours she dedicated to discuss and read my research.

I thank Roger Levy, Joakim Nivre, and Laura Rimell, who kindly agreed to be members of the defence committee, and Ur Shlonsky for serving as the president of the jury.

The research visits to Paris and Edinburgh were highlights of the years of my doctoral studies. I would like to thank Benoit Crabbé and Frank Keller for hosting me at their institutions and helping me develop my research in new exciting ways. I would like to acknowledge the funding by Labex EFL and Swiss National Science Foundation which provided me with these opportunities.

I am grateful to James Henderson for leading the CLCL group, and all my colleagues at CLCL and the department who were there for me in my everyday office life. I would like to say special thanks to Tanja Samardžić, Sarah Ouwayda, Majid Yazdani, Nikhil Garg, Sharid Loáiciga, Yves Scherrer, Corentin Ribeyre, Alexandre Kabbach, and Hasmik Jivanyan, who have become my friends over the years of shared experiences and hard work.

This thesis would not be written without Lorenzo and the beauty of nature by my side. Their existence alone will take me through everything.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Word order is one of the most readily observable parts of the grammatical system of human languages. Contrasts between the grammatical (1.1a) and the ungrammatical (1.1b) sentences, which differ minimally in their word order, constitute linguistic facts that a syntactic theory seeks to explain.

(1.1) a. John ate a cake with a fork
b. *John ate with a fork a cake

By *word order*, we understand the order between lexical elements and phrases in a syntactic relation, e.g., a verb-object relation *ate* → *cake* or a verb-modifier relation *ate* → [*with a fork*]. Syntactic relations form the hierarchical tree structure of an utterance. This tree structure is mapped onto a one-dimensional sequence of sounds or written symbols resulting in the observed word order.

Languages of the world vary greatly in the word order constraints specified in their grammar. For instance, the main syntactic elements of a clause — verb (V), subject (S) and object (O) — are arranged in the SOV order in Japanese (1.2), as opposed to English, which places them in the SVO order.

(1.2) John-ga keiki-o tabeta
John-*nom* cake-*acc* ate

'John ate a cake'

An even more intriguing observation is that word order constraints can be more or less flexible in different languages. The parallel sentences in English, Italian and Russian illustrate three constructions of varying word order flexibility (1.3–1.5).

(1.3) a. *English*: I saw Mary in the shop / *I saw in the shop Mary
b. *Italian*: Ho visto Maria al negozio / *Ho visto al negozio Maria
c. *Russian*: Я видел Марию в магазине / Я видел в магазине Марию

(1.4) a. *English*: I saw that John came / * I saw that came John
b. *Italian*: Ho visto che Gianni è venuto / Ho visto che è venuto Gianni
c. *Russian*: Я видел что Иван пришел / Я видел что пришел Иван

(1.5) a. *English*: I saw a new book / *I saw a book new
b. *Italian*: Ho visto un nuovo libro / Ho visto un libro nuovo
c. *Russian*: Я видел новую книгу / *Я видел книгу новую

In example (1.3), English and Italian require the direct object *Maria* to be adjacent to the verb *saw* while Russian also allows it to appear after the prepositional phrase *in the shop*. In (1.4), the SV order *John came* is the only grammatical option in English but it can be reversed in the embedded clause in Italian and Russian (*è venuto Gianni*, *пришел Иван*). Example (1.5) shows that English and Russian have only one grammatical position for adjective modifiers which must appear before the noun. In Italian, both pre-nominal and post-nominal positions are allowed.

The cases of availability of two orders with equivalent semantic meanings as in examples (1.3c), (1.4b,c) and (1.5b) are known as cases of *word order variation*. Such cases are challenging for a syntactic theory because they imply that one underlying syntactic structure can be mapped onto several word order realisations. Word order variation phenomena are not explicable with a one-to-one structure–linearisation mapping and are not purely syntactic. Rather, these phenomena emerge at the interface between the syntactic and the production systems of the language. The preferences between two alternative grammatical orders and the choices observed

in production data are affected by various types of factors, including processing constraints and discourse context.

The linguistic data illustrated in examples (1.3–1.5) raise two related fundamental questions. The first question concerns the variation observed across languages: To what extent do languages vary in their word order and its flexibility, and why? Describing word order constraints and identifying the limits of their variation is important for the typological study of languages. Explaining why these limits exist leads to an improved understanding of the universal properties of languages and their structures.

The second question is concerned with the variation observed in an individual language, e.g., in Italian in examples (1.4b) and (1.5b) and in Russian in examples (1.3c) and (1.4c). When two word order options are available, why is one option chosen over the other option in a given sentence? This question can be answered thoroughly only by looking at many factors: both the ones general to a language (Italian or Russian) and the ones specific to the context of speech. By analysing intra-linguistic word order variation in a cross-linguistic perspective, we can further identify the factors which are general not only to one language but to all languages.

This dissertation aims to advance our understanding of word order variation from these two perspectives. Our focus is, in particular, on one common characteristic of word order in natural languages: the tendency for related syntactic elements to appear close to each other in the linearisation of the structure.

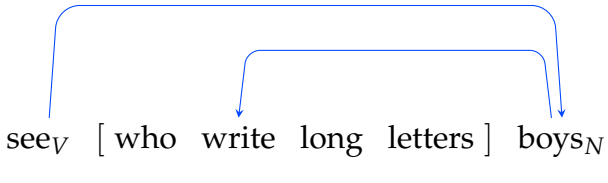
1.1 Dependency length minimisation

It has repeatedly been observed that languages tend to minimise the distance between words and phrases connected by a syntactic relation. In examples (1.1) and (1.3a) in English, the verb (head) and the direct object (its syntactic child) must be linearly adjacent: the order V PP O with an intervening prepositional phrase (PP) is not grammatical. This tendency is found both across grammatical orders of languages of the world and in intra-linguistic word order variation patterns.

Typological data tell us that word order patterns which produce short distances are cross-linguistically more frequent than patterns which produce long distances (Greenberg, 1963; Hawkins, 1994; Dryer, 1992). For example, some of the frequently attested orders between a verb, its nominal object (N) and a relative clause modifying the object (RelC) are V [N RelC] order (1.6) (e.g., as in English) and [RelC N] V order (1.7) (e.g., as in Japanese). By contrast, the order V [RelC N] (1.8) is rarely found (Dryer and Haspelmath, 2011).[1]

(1.6) 

(1.7) 

(1.8) 

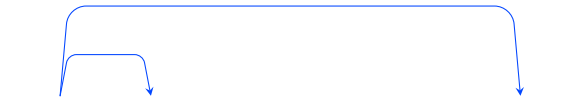
It is evident from the schematic representation of the distances between syntactically-related elements (*see* → *boys*, *boys* → *write*) that the orders (1.6) and (1.7) place these elements closely adjacent to each other while the order (1.8) places them further apart.

Similar tendencies have been extensively observed for cases of word order variation involving a choice between two or more possible grammatical orders in a language. It was demonstrated using corpus and experimental production data that speakers

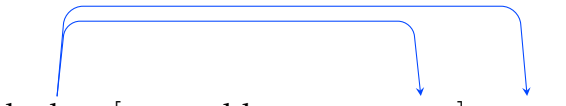
[1] According to World Atlas of Language Structures (`http://wals.info/combinations/83A_90A`), the first two orders are attested in 415 and 132 languages, while the third one is found only in five languages of the world.

use more frequently the order which yields smaller distances between dependent words compared to alternative linearisations (Hawkins, 1994; Stallings et al., 1998; Wasow, 2002; Gries, 2003; Bresnan et al., 2007). Consider, for instance, the case of word order variation involving a phrasal verb with particle and a nominal object phrase in English (*look*$_V$ *up*$_{Prt}$ [*a story*]$_{NP}$). The order V Prt NP (1.9) was found to be preferred compared to the order V NP Prt (1.10) when the object noun phrase is long.

(1.9)



(1.10)



As can be seen from the illustrations in (1.9–1.10), the distances between V and Prt (*look* → *up*) and V and N (*look* → *story*) are shorter in the preferred order V Prt NP.

Following recent work (Temperley, 2007; Park and Levy, 2009; Tily, 2010; Futrell et al., 2015b), we refer to these tendencies in typological distributions and word order variation preferences cumulatively as a *dependency length minimisation (DLM)* principle. The arcs in examples (1.6–1.10) indicate syntactic *dependencies* between words and *dependency length* is the distance between two dependent words.[2]

DLM emerges as a universal bias affecting word order, manifested in typological distributions and phenomena of word order variation in many languages. Despite the fact that the DLM principle has already received much attention, there remain many open fundamental and intriguing questions. To what extent is DLM universal?

[2]Dependency length is measured as a linear distance between words. Note, however, that this notion relies on the hierarchical representation of the sentence. For example, in the phrase *write a note with a pencil* we compute the dependency length between the modifier *with a pencil* and its head *write*, not between the modifier and a potential head *node* which is linearly adjacent.

The cross-linguistic studies of Liu (2008), Gildea and Temperley (2010) and Futrell et al. (2015a) found that all investigated languages are shaped by DLM (on average, across all constructions in a language). However, some of the languages including German and SOV languages such as Persian and Japanese have longer dependencies than other languages. Further investigations are required to understand how the effect of DLM differs between languages and why. If DLM is a universal pressure, this also suggests that it should apply to all cases of word order variation. Previous work found DLM effects in many word order variation constructions, but mostly of one structural type including alternation of only two dependencies as in (1.9–1.10). It is an open question whether DLM indeed affects all types of variation and how the interaction with other factors influences it. Investigating these aspects of DLM should help us answer the most perplexing question: What is the nature of the DLM principle? Does it originate in constraints on production or comprehension processing mechanisms or is it a more general communication pressure?

The goal of this thesis is to provide new empirical facts and theoretical considerations towards answering these questions. We extend previous work based on syntactically-annotated cross-linguistic corpus data by analysing and modelling word order variation and DLM at three different linguistic levels: on average in a language, in individual word order variation constructions, and in the online mechanism of sentence linearisation.

1.2 Computational analysis of word order

This thesis pursues an empirical approach to linguistic theory: we study fundamental questions about word order using corpus data and computational modelling.

Quantitative analysis of word order in a language requires specialised corpus data. To establish statistical properties of word order patterns we need a large enough sample of naturally produced sentences. Written corpora are the most common source of such data. Nowadays, we can gather vast amounts of written digital text. The raw text data are not, however, sufficient to analyse word order variation: we need to know the syntactic structures of the sentences which are crucial for

investigating word order phenomena as a mapping between syntactic structure and its linearisation. The requirement to have access to the syntactic analysis of a sentence ties our empirical investigations to syntactically-annotated corpora, known as treebanks. The experiments presented in this thesis rely on certain properties of the treebanks and the choices of syntactic annotation they make. In particular, the treebanks that we employ provide dependency grammar annotation of the sentences which define the way we compute dependency lengths.

Syntactic treebanks require manual annotation by linguistic experts and are therefore expensive and scarce linguistic resources. As a consequence, until recently, quantitative analyses of word order variation based on syntactic treebanks were habitually conducted on one language of choice, commonly English. Fortunately, the number of treebanks available in different languages grows every year. Moreover, since treebanks serve as essential training and evaluation resources for natural language processing applications, a systematic effort of the community is directed towards harmonising and unifying treebanks and their annotation designs (Zeman et al., 2012; Petrov et al., 2012; de Marneffe et al., 2014; Nivre et al., 2016). The availability of collections of treebanks with the same syntactic annotation such as the Universal Dependencies treebanks (Nivre et al., 2016, 2017) provides opportunities to perform large-scale analyses of syntactic phenomena in cross-linguistic perspective. The experiments in this thesis leverage these new linguistic resources and analyse a total of 15 languages, contributing to the recent line of large-scale cross-linguistic research on word order.

We use computational modelling of empirical data as a means to investigate and formally test theoretical linguistic hypotheses. The primary type of observations that come from our data is the frequencies of occurrences of word order options. One way to establish the preferences in word order variation constructions is simply by comparing the frequency of two word order options (e.g., V Prt NP and V NP Prt) in a corpus. Of course, the choice between two possible linearisations is subject to many factors and constraints. We are primarily interested in teasing apart and testing the effects of these factors. To do so, we use logistic regression statistical models which have been traditionally applied in corpus analyses of syntactic variation (Gries, 2001; Bresnan et al., 2007). These models allow us to examine many factors potentially affecting variation and measure their effects.

The drawback of logistic regression models is that they simplify the variation phenomena by focusing only at surface factors observed in a sentence and its structure and abstract away from complex interactions between the factors, the discourse context of the utterance and the mechanisms of language production and comprehension. To address some of these limitations, we develop a computational model which is designed to model the process of word order production explicitly. This model can be seen as an implementation of one stage of the online language production mechanism: the mapping of the syntactic structure onto the order of words. This model is a type of machine learning model. It is trained on linearised dependency trees provided by a treebank to predict the next word in an utterance given the previously produced words and the rest of the dependency tree. This second type of computational modelling provides a means to test how the constraints on the language production mechanism, such as online processing and memory limitations, affect word order variation.

The computational methods applied in this thesis are connected in several ways to natural language processing (NLP) research. While the focus of our work is on linguistic questions concerning word order variation, some of the methods we develop derive from the NLP models for text processing and language generation, more specifically, statistical parsing and surface realisation models. Conversely, we also find that the models and quantitative analyses we develop in this work are relevant and useful for NLP research.

1.3 Overview of the thesis and its goals

The work presented in this thesis investigates cross-linguistic word order variation phenomena related to dependency length minimisation through the use of syntactically-annotated data and computational models. Three experiments which we present in Chapters 3 through 5 address word order variation phenomena at three different linguistic levels: the language level, the construction level and the sentence level. In principle, these experiments and their results can be viewed as independent

pieces of work, but together they aim to provide a new unified perspective on word order variation and DLM across different linguistic levels.

Before presenting our experiments and results, we outline the general theoretical and methodological context for our work in Chapter 2. We start by describing the treebank data we use. As mentioned previously, we use dependency-annotated treebanks. We highlight the main properties of this grammatical annotation and the reasons for adopting it in our work. Next, we discuss the theoretical syntactic assumptions on word order which we implicitly adhere to when we analyse word order as a mapping between dependency trees and their linearisations. We present previous quantitative corpus-based work on word order variation underlining the empirical methodology we follow in our work. As part of this chapter, we review a large part of the work on DLM focusing on the evidence for DLM effects in word order variation constructions as well as the main processing explanations of DLM.

Chapter 3 presents our first study dedicated to the typological language-level investigation of word order properties. We analyse word order in a language as a whole with a goal to measure and compare the degree of dependency length minimisation and word order freedom across languages. Previous work has shown statistically that many typologically-different languages tend to minimise dependencies at language level (Futrell et al., 2015a). The extent of this minimisation is, however, varied: some languages like English or French seem to minimise dependencies more than other languages such as Persian or Russian. Is it possible to quantitatively compare the rate of DLM across languages? We show that the answer is yes, by using statistics extracted from dependency treebanks. We analyse previously proposed measures of dependency lengths and show that we can robustly compute the degree of DLM in a language based on dependency length of a sentence relative to the minimal possible dependency length. A related question we are interested in is whether we can compare the degree of word order freedom across languages in a similar way. Word order variation at the level of language can also be computed using dependency treebanks, but it is harder to do in a statistically-robust way. Language-level measures of word order properties can contribute not only to linguistic typology but also to NLP research. To illustrate their potential application for NLP, we evaluate the measures of DLM and word order freedom as the correlates of statistical parsing

performance and show that they can be applied to diagnose and inform parsing systems.

Chapter 4 aims to investigate the relevance of the general DLM principle, as used at the treebank level in Chapter 3, in word order variation in one syntactic construction. We focus on adjective-noun variation in Romance languages. Many adjectives can appear both before the noun and after the noun in Italian (see, e.g., example (1.5b)) and other Romance languages. This alternation has received substantial attention in theoretical, empirical and computational linguistic studies. However, previous research focuses mainly on the semantic and lexical constraints of the prenominal and postnominal adjectival positions. Apart from the analysis of heavy adjective phrases (Abeillé and Godard, 2000), this construction has not been previously investigated in connection with DLM. Consequently, we use the adjective-noun word order variation to probe the universality of the DLM principle on a new syntactic phenomenon. We start by formalising the general cumulative DLM principle for this construction (Temperley, 2007). Adjective-noun variation is structurally different from alternations such as verb-particle shift (examples (1.9–1.10)) used traditionally as evidence for DLM, and the predictions of DLM are not straightforward since they depend on the composition of the whole noun phrase. To verify these predictions, we conduct several systematic corpus-based statistical analyses in five Romance languages. To our knowledge, this is the first large-scale corpus study of adjective placement across several Romance languages. Our results reveal several types of DLM effects in complex noun phrases with adjectives. Interestingly, some of the DLM-induced adjective distribution patterns have not been reported in the previous literature.

The experiments in Chapters 3 and 4 focus on word order distributions observed at two different linguistic levels: in a language as a whole, and in one specific syntactic construction and its realisations, respectively. Apart from assuming common DLM effects, these studies analyse two types of word order distributions independently, without referring to the shared processes of language production which generated these distributions. In Chapter 5, we pave the way for studying word order distributions at different linguistic levels using one integrated approach: by modelling and analysing the word order production system directly.

The word order part of the language production, which we focus on, is known as sentence linearization. Our proposed model of sentence linearization is trained to learn the mapping between the syntactic structure (unordered dependency tree) and the word order in a sentence. Crucially, it is designed to produce word order in an online fashion, that is, word-by-word. The choice of the next word is based on the acquired probabilistic grammar and is made greedily with minimal use of computational and memory resources. These architectural features make our system a cognitively plausible implementation of an incremental word order production process. Moreover, the choices between alternative grammatical word orders can be naturally integrated into this system as a re-ranking step. Making a connection with Chapter 3, we confirm the language-level DLM in a new way through the modelling of word order based on local word order production decisions.

1.4 Publications

Part of the work presented in this thesis was previously published as the following peer-reviewed papers.

Chapter 3 is largely based on the two papers:

- Kristina Gulordava, Paola Merlo (2015a) *Diachronic Trends in Word Order Freedom and Dependency Length in Dependency-Annotated Corpora of Latin and Ancient Greek.* International Conference on Dependency Linguistics

- Kristina Gulordava, Paola Merlo (2016) *Multi-lingual Dependency Parsing Evaluation: a Large-scale Analysis of Word Order Properties using Artificial Data.* Transactions of ACL

Chapter 4 draws on the work published in:

- Kristina Gulordava, Paola Merlo, Benoit Crabbé (2015) *Dependency length minimisation effects in short spans: a large-scale analysis of adjective placement in complex noun phrases.* Proceedings of ACL

- Kristina Gulordava, Paola Merlo (2015b) *Structural and lexical factors in adjective placement in complex noun phrases across Romance languages.* Proceedings of CONLL

Chapter 2

Background

The work presented in this thesis draws on many adjacent research fields interested in phenomena of word order and dependency length minimisation: theoretical and empirical syntax, psycholinguistics, natural language processing, typology. This chapter provides a general high-level picture of the previous work on the main two topics of the thesis. At the same time, we make explicit our main starting assumptions about the grammatical structure and its relation to word order. We focus our discussion on the methodology and findings of the previous corpus-based word order variation studies addressing DLM effects. In the following chapters, we discuss in more detail the previous work which is related specifically to each of the three experiments.

We start this chapter by first describing the dependency grammar formalism and dependency treebank resources which constitute the empirical basis for our work.

2.1 Dependency treebanks

A treebank is a text corpus where each sentence is annotated with its syntactic structure. The release in 1994 of the English Penn Treebank (Marcus et al., 1994) — the first large-scale treebank of 1.6 million words — paved the way for empirical

syntactic analyses and the development of data-driven techniques to automatic natural language parsing.

Dependency treebanks are treebanks which annotate dependency syntactic structure of sentences, as opposed to phrase structure (used, for example, in the Penn Treebank annotation). The first large-scale dependency treebank is the Prague Treebank. Its development was inspired by the long-standing Praguian linguistic tradition and the theory of the dependency-based Functional Generative Description (Hajičová, 1998; Böhmová et al., 2003). Recently, there has been a growing interest in developing dependency treebanks for many languages. The consolidation effort around building and harmonising dependency treebanks has lead to the Universal Dependencies initiative (McDonald et al., 2013; de Marneffe et al., 2014; Nivre et al., 2016) which to this moment (v2.1, 15 November 2017 release, Nivre et al. (2017)) has produced 102 treebanks in 60 languages.

The availability of multilingual syntactically-annotated corpora in the form of dependency treebanks creates a starting point for the work presented in this thesis. This section describes the main syntactic principles underlying dependency structure analyses and the properties of the treebanks which are essential for the experiments described in the following chapters.

2.1.1 Dependency structure representation

The origins of the dependency grammar tradition date back to the work of Lucien Tesnière (Tesnière, 1959, 2015). Some of the most developed dependency grammar frameworks are the Prague School's Functional Generative Description (Sgall et al., 1986), Mel'čuk's Meaning-Text Theory (Mel'čuk, 1988), and Hudson's Word Grammar (Hudson, 1984).

All the types of dependency grammars share the following main properties. First, adopting the lexicalist hypothesis in syntax, the units of the syntactic structure are assumed to be words. Secondly, the words are connected by binary grammatical relations called *dependencies*. A dependency $w_1 \rightarrow w_2$ between two words w_1 and w_2 is always asymmetric. It states that the word w_2 is dependent on the word w_1 which

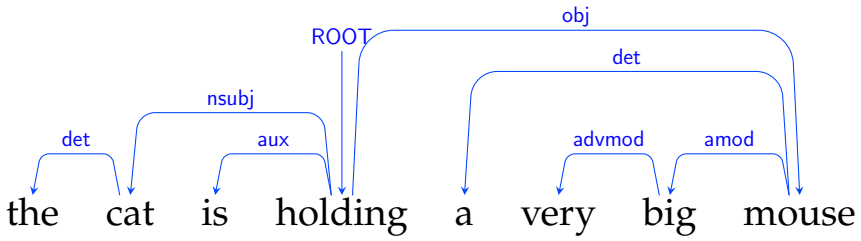


Figure 2.1: An example dependency tree of the English sentence *the cat is holding a very big mouse*.

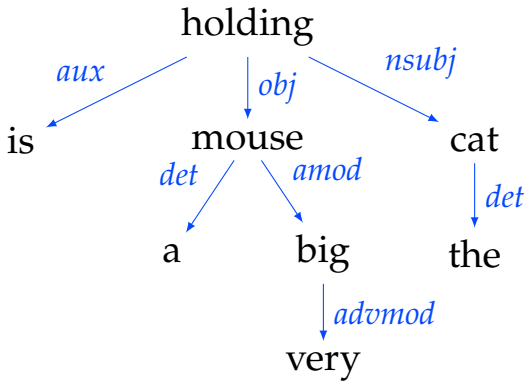


Figure 2.2: An unordered dependency tree representation of the English sentence *the cat is holding a very big mouse*.

is called the *head* of the dependency. The dependency relations between the words in a sentence form a tree structure as illustrated in the Figure 2.1 for a simple English sentence. Note that the dependencies are represented by arrows pointing from the head to the dependent.

The tree condition implies that each word — node in the tree — has one and only one head. The root word of a sentence is indicated by a special dependency, e.g., ROOT $\rightarrow$ *holding* in our example.

In addition to the tree structure composed of binary dependencies, dependency grammars specify a set of labels to distinguish between different types of dependencies. For example, the label *nsubj* indicates that *cat* is the subject of the verb *holding*, while *mouse* is the object of the verb identified by the label *obj* (Figure 2.1).

Importantly, a dependency analysis indicates hierarchical relations between lexical

units (words) and gives, in principle, a syntactic representation which is independent of the order of words in a sentence. Figure 2.2 illustrates a hierarchical dependency representation of the sentence, which we call *unordered dependency tree* as opposed to *ordered dependency tree* in Figure 2.1.

The fact that a dependency structure is assigned using a small set of criteria for identifying the head-dependency relations and that these criteria do not depend on the contiguous sequences of words (as in phrase-structure analyses) makes it a natural choice for annotation of languages with relatively free word order. Consider an extreme example of non-configurational sentence structure in Latin (Figure 2.3). The dependency structure annotation of this sentence using the available morphological information is straightforward despite the non-contiguous word order.[1] This example also points to why dependency representation is advantageous for annotation of languages with different morphosyntactic properties under one annotation scheme. If we take a sentence similar to the one in Latin in some other language, say in English, the (unordered) dependency structure could be given in a very similar way for the two languages, despite some crucial differences in the word order.

These reasons, among others, have led to the gradual adoption of dependency-structure annotation for building new treebanks and to a recent effort to unify the existing dependency annotation schemes under one annotation scheme, known as Universal Dependencies (Nivre et al., 2016). In this work, we used Universal Dependencies treebanks for most of our experiments with addition of several treebanks from the PROIEL project Haug and Jøhndal (2008). We discuss these treebanks and the Universal Dependencies annotation scheme in the next section.

2.1.2 Dependency treebanks and annotation schemes

Treebanks come in different forms. Apart from the annotation of syntactic dependencies, a typical treebank contains other levels of linguistic information. This

[1] This example in Latin also illustrates so-called *non-projective* dependencies. An arc between two words is defined as non-projective in an ordered dependency tree if among the words it spans there is a word belonging to a different subtree. A presence of a non-projective arc corresponds to non-contiguous constituents in the phrase structure of the sentence.

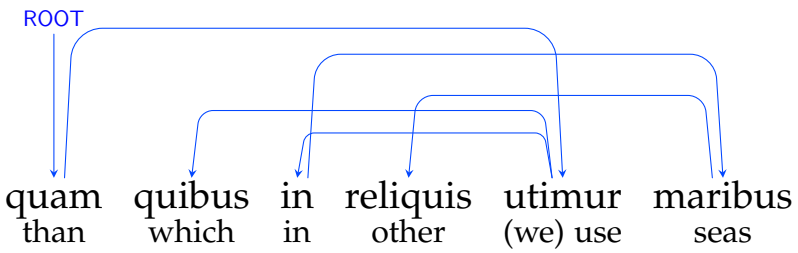


Figure 2.3: The dependency tree of a phrase in Latin, extracted from the Caesar PROIEL treebank (Haug and Jøhndal, 2008) translated as *than those which we use in other seas*.

information most frequently covers morphosyntactic properties at the level of the word. The most common and useful word features are *lemmas*, *part-of-speech* tags (PoS tags) and morphological features (case, number, gender, tense and other).

A popular textual format for dependency treebanks, known as CONLL format, represents words as lines and features as columns. Each word has an index indicating its linear position in the sentence. The dependency tree structure is provided by indicating for each word w the index of its head node h. The dependency label for a $h \rightarrow w$ relation is similarly indicated as a feature of the dependent word w. This simple format allows extracting dependencies of a specific type using simple data processing scripts. This format is widely adopted and helps to process different treebanks and languages in the same manner.

A bigger challenge for automatic processing and analysis of multilingual dependency treebanks are conceptual differences in the annotation decisions. Treebanks can differ in the set of PoS tags or dependency labels that are used for annotation or, perhaps more crucially, in the criteria to choose heads and attach dependents. In recent years, there has been a substantial effort towards unifying and harmonising treebanks at all levels of annotation. The first step concerned the design of a small set of coarse-grained Universal PoS tags (Petrov et al., 2012).[2] Many NLP systems rely on PoS tagging to provide higher-level syntactic or semantic analysis of sentences. The diversity of PoS annotation (ranging from 11 to 294 language-specific tags in 25

[2]See also Buchholz and Marsi (2006); Rambow et al. (2006); Nivre et al. (2007) for previous work in this direction.

treebanks analysed by Petrov et al. (2012)) prevents the off-the-shelf application of systems developed in English to other languages. Based on these practical considerations, Petrov et al. (2012) proposed and evaluated a set of 12 coarse universal PoS categories. They provided the mapping of the tag sets of 25 existing treebanks to the universal PoS tags and highlighted the resulting practical advantages for NLP applications. For instance, they demonstrated that the transfer of parsing systems from one language to another improves substantially when the languages share the same universal PoS tagset.

At the moment, the Universal PoS tag set comprises 17 categories including content word open class categories: NOUN, PROPN (proper noun), VERB, ADJ (adjective); function word categories: PART (particle), ADP (adposition, i.e., both pre- and post-positions), AUX (auxiliary), DET (determiner), NUM (numeral), PRON (pronoun), SCONJ and CCONJ (subordinate and coordinate conjunctions); and additional categories such as PUNCT (punctuation) and SYM (symbol).[3]

HamleDT treebank collection (Zeman et al., 2012) is the first large-scale attempt to harmonise the annotation schemes of existing dependency treebanks. It features, in particular, the same set of dependency labels across all treebanks in the collection which were automatically mapped from the language-specific dependency labels. The HamleDT annotation is based on the Prague Dependency Treebank scheme.

These initiatives have contributed to the most recent and large-scale collaborative effort carried out by the computational linguistic community and known as Universal Dependencies (UD) project.[4] It is aimed at designing a dependency structure annotation scheme which can be applied to typologically diverse languages. As for Petrov et al. (2012), one of the motivations behind UD project is the need to improve cross-linguistic transferability and comparison of automatic NLP systems. This ambitious goal is achieved by trading off detailed syntactic analyses provided by language-specific annotation choices for simplification of the syntactic structures and dependency relations. This unification and simplification necessarily starts at the level of parts-of-speech which are annotated using the coarse Universal PoS tag set.

[3] http://universaldependencies.org/u/pos/
[4] http://universaldependencies.org

In this thesis, we took advantage of the availability of dependency treebanks for many languages and, in particular, the many treebanks annotated by the Universal Dependencies project. We rely on the specific properties of the UD annotation scheme in two ways. First, we access automatically a set of relevant syntactic constructions (such as noun phrases containing an adjective in the experiments in Chapter 4) by matching PoS tags (ADJ, NOUN) and dependency labels (amod). Secondly, we compute the lengths of dependencies which are by definition conditioned on the form of dependency trees. The linguistic decisions of the annotation scheme play, therefore, an essential role in the interpretation of the results of our experiments. We describe the details of the Universal Dependencies annotation scheme, which we used in the majority of our experiments.

Universal Dependencies annotation scheme

A distinctive property of the UD annotation is its adherence to the *content-head* principle of assignment of dependencies. The annotation builds on the Stanford Dependencies scheme (de Marneffe and Manning, 2008) which was proposed as a description of grammatical relations (initially for English) with an emphasis on interpretable, "semantically contentful" relations useful for down-stream NLP applications such as information extraction or question answering. These applications aim to recover the meaning of sentences and texts which often depends crucially on non-local syntactic relations. To do this, such applications rely on automatic syntactic parsing in their pipeline. As a consequence, the Stanford Dependencies were designed to favour the direct extraction of predicate-argument structure and other relations between content words. This is in contrast with traditional syntactic formalisms which give much prominence to function words. Stanford Dependencies, for example, make the choice of treating determiners and auxiliaries as modifiers. However, prepositions are still considered the heads of the prepositional phrase. The last version of the Stanford Dependencies scheme (de Marneffe et al., 2014), which eventually developed into the current UD scheme, consistently chooses content words as heads in a dependency tree. The main motivation for this convergence is the cross-linguistic parallelism. If dependency relations are drawn between content

words then they will hold also in languages without some particular function words (e.g. prepositions or determiners).

The content-head UD representation is illustrated in the example in Figure 2.4. The content words *staring*, *cat*, *mouse* are connected using fundamental grammatical relations such as *nsubj* and *iobj*. The function words *is* and *a* are dependents and modifiers of the content words. In contrast, Figure 2.5 shows a function-head annotation variant of a similar sentence. Here, for example, the auxiliary *is* is chosen to be the head of the subject on the syntactic grounds that there is an agreement relation aligning the number of the subject and the number of the auxiliary *be*. Similarly, preposition *at* is chosen to be the head of the prepositional phrase because the verb *stare* is in a direct relationship with the preposition *at* which it is said to select (the verb cannot appear with some other preposition, e.g., *on*). In contrast, in the content-head annotation, prepositions are modifiers of the nouns which are heads of prepositional phrases (e.g., *mouse* is the head of PP *at a mouse*).[5]

Using the phrase-structure terminology, the basic head-assignment decisions in the UD annotation scheme can be summarised as follows. The head of the noun phrase (NP) is a noun, all the noun dependents such as determiners, numerals, adjectives are its modifiers. The head of the prepositional phrase [P NP] is the head noun in the NP, and the preposition is its dependent. The main content verb is the head of the verb phrase, and all the auxiliary verbs (modal, tense and aspect modifiers) are its dependents.

The advantages of the content-head annotation are evident in a cross-linguistic comparison. Consider the sentence in English from Figure 2.4 translated into Russian (Figure 2.6). Importantly, the shared predicate-argument structure of the sentence (*cat* $\leftarrow$ *staring* $\rightarrow$ *mouse*) is annotated by the same dependencies in the English sentence and in the Russian sentence. On the other hand, additional syntactic information is expressed differently in two languages: Russian uses morphological case marking to

[5]The content-head choices of the UD annotation scheme are controversial from a syntactic stand, as seen, for example, from the critical take of Osborne (2015). In fact, UD is not proposed as a syntactic theory (Nivre, 2015, p. 3). Rather, dependency relations serve to capture both surface syntactic relations and deeper predicate-argument relations situated on the interface between syntax and semantics.

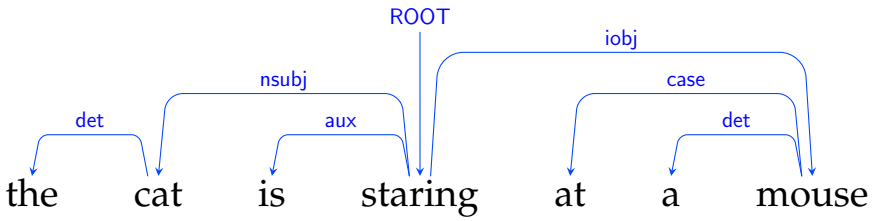


Figure 2.4: The content-head dependency tree of the English sentence *the cat is staring at a mouse*.

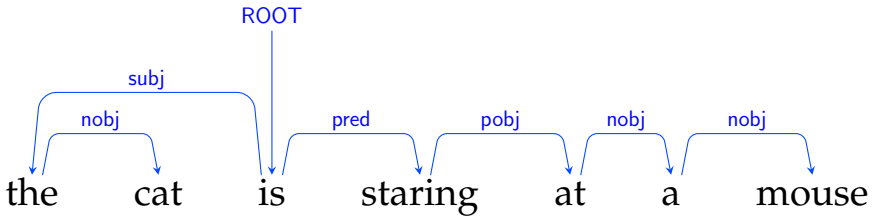


Figure 2.5: The function-head dependency tree of the English sentence *the cat is staring at a mouse*.

express grammatical relations while English utilises function words — prepositions. The distinction between core content word relations and secondary relations between function and content words allows drawing parallel structural analyses between syntactically different languages, which would not otherwise be possible.

The UD dependency labels comprise 35 categories which distinguish between core and non-core predicate dependencies as well as nominal versus clausal dependencies. For example, *nsubj* and *csubj* annotate nominal and clausal subject relations respectively, while *obj* and *iobj* annotate direct object and indirect nominal objects. The UD

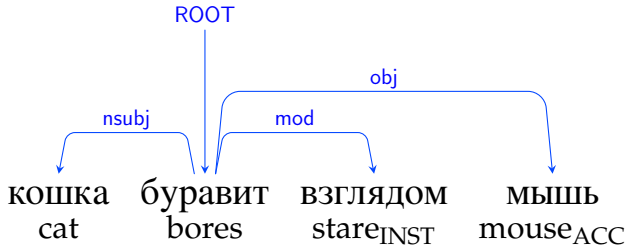


Figure 2.6: The Russian phrase corresponding to the English *the cat is staring at the mouse*.

documentation contains extensive reference on all relation types and many example analyses of constructions.[6]

UD treebanks

At the moment, the latest version of UD treebanks (v2.1) released in November 2017 counts 60 languages and 102 treebanks. Among these, 30 languages have more than 100.000 annotated tokens. The work presented in this thesis was conducted using the UD treebanks released with versions 1.2 and 1.3 which cover between 33 and 40 different languages. We provide the statistics of the subset of the treebanks for each experiment separately.

PROIEL treebanks

In addition to UD treebanks, we used treebanks annotated as part of the PROIEL project (Haug and Jøhndal, 2008). PROIEL project provides annotation for literary texts in ancient languages such as Latin, Ancient Greek, Old Church Slavonic, Classic Armenian and many others. We use the treebanks of Latin and Ancient Greek in our experiments presented in Chapter 4. Starting from the UD release v2.0, the PROIEL treebanks were converted to the UD annotation scheme and included in the collection. However, we used the original PROIEL treebanks and not their UD versions because we analyse and compare different literary works. The UD version contains only one, composed, treebank per language. Instead, on the PROIEL website, the treebanks of Latin comprising the works of Caesar, Cicero, Vulgate's Bible and others can be downloaded separately.[7]

The PROIEL annotation scheme is based on the Prague Dependency Treebank scheme and is similar to the UD scheme. There are some minor differences in the naming of dependency labels and one structural difference in the head-assignment rules: i.e., the prepositions are considered to be heads of prepositional phrases. However, these

[6] http://universaldependencies.org/u/dep/index.html
[7] https://proiel.github.io/

properties do not make a difference for our analyses, and we leave the reader to consult the PROIEL annotation guidelines for further information.[8]

2.2 Theoretical and empirical framework

Constraints on word order are part of the grammatical knowledge of speakers. A speaker of English knows that (2.1a) is an acceptable, grammatical sentence in English while (2.1b), with a slightly different word order, is ungrammatical:

(2.1) a. John ate a cake with a fork
b. *John ate with a fork a cake

Yet, syntactic theories differ substantially in how much prominence they attribute to word order and where the word order constraints are placed in the grammatical representation. In the mainstream phrase-structure transformational syntax starting from Chomsky (1957), word order is tightly connected to the hierarchical grammatical representation. First, it is specified by grammar as part of phrase-structure rules creating the "deep" structure. It is then subsequently modified by movement transformations resulting in "surface" structure (Chomsky, 1965; Kayne, 1994; Rizzi, 2004; Cinque, 2005). In a most straightforward analysis of this kind, a phrase-structure rule VP $\rightarrow$ V NP PP captures the word order pattern in the example (2.1). A more elaborate account proposed in Principles and Parameters and Government and Binding framework (Chomsky, 1981; Chomsky and Lasnik, 1993) defines a small set of general linearisation principles applied to the X-bar phrase structure. These rules fix the order between all the specifiers and their heads and all the complements and their heads. A parameter such as "complements appear to the right of their heads" captures many regularities in one language (in English object appears on the right of the verb; relative clauses on the right of their complementisers; PP complements on the right of the noun). Varying the value of this parameter captures the typological variation between head-initial and head-final languages (e.g., in Japanese the complements appear to the left of their head; the reverse of the previous statements for English is

[8] http://folk.uio.no/daghaug/syntactic_guidelines.pdf

true). Adjuncts are treated differently compared to arguments, and their position is assumed to be underspecified with respect to the head by the grammar.[9]

In non-transformational syntactic formalisms, in particular, generalised phrase structure grammar (Gazdar et al., 1985) and head-driven phrase structure grammar (Pollard and Sag, 1994), the treatment of word order is conceptually different. Since there are no movement operations modifying "deep" linearisation, word order can be separated more explicitly from the structure-building part of the grammar. A body of work proposed, more specifically, that *immediate dominance* relations should be decoupled from *linear precedence* relations (Pullum, 1982; Uszkoreit, 1983; Falk, 1983; Reape, 1993).[10] Instead of using one phrase-structure rule to capture the data in 2.1, an immediate dominance rule VP $\rightarrow$ { V, NP, PP } could state that a verb phrase consists of a verb, a noun phrase and a prepositional phrase, and a linear precedence (word order) rule NP $<$ PP could indicate that an object noun phrase precedes an adjunct prepositional phrase. The use of decoupled representations in this work is motivated as a means to address free word order variation phenomena such as scrambling in German. It is sufficient not to specify the precedence between two phrases to allow formally for both orders (NP PP and PP NP).

The dependency grammar tradition adheres to a similar distinction between hierarchy and word order, where the immediate dominance relations are known as the *tectogrammatical* layer of syntactic analysis (Sgall et al. (1986), see also Dowty (1996)). As we have seen in the previous section, this layer of analysis is given by an unordered dependency tree (Figure 2.2). In other words, immediate dominance phrase structure rules are equivalent to dependency relations of type V $\rightarrow$ N, V $\rightarrow$ P, without intermediate phrase structure nodes.[11]

[9] A more recent Minimalist approach to syntax (Chomsky, 1995) challenges the status of word order as part of the grammar, e.g., specified by phrase-structure rules, movement or language-level parameters. The main structure-building operation *Merge* which combines two syntactic phrases X and Y is assumed to determine the hierarchical relationship between the two (e.g., whether X or Y is the head) but does not specify the order between them. For Chomsky, word order is part of the spell-out process producing phonological form of utterances, but not the syntax proper.

[10] Some earlier work in the transformational grammar tradition proposed analyses similar in spirit to account for non-configurational languages and word order variation (Hudson, 1979; Hale, 1983).

[11] Dependency trees have a more shallow structure (with fewer depth levels and more children under one head) than traditional phrase structure trees. Consequently, there is, in principle, more combinatorial freedom to specify linear precedence rules in dependency grammars.

In this work, our focus is exclusively on the linear precedence relations, and we treat the dominance layer of syntactic analysis as given by the unordered dependency trees annotated in the treebanks. In the context of the cross-linguistic analysis, we assume that hierarchical syntactic structure is comparable across typologically different languages with different word order properties. As we argued in Section 2.1, the dependency grammar analyses provided by the Universal Dependencies annotation are built on the same assumption.

The two-level organisation of the grammar — decoupling hierachical from linear information — is well supported by psycholinguistic models of sentence production (Bock and Levelt, 1994). The experimental evidence for this distinction comes from speech errors such as substitution or agreement errors (see Ferreira and Engelhardt (2006) for an overview). We discuss these arguments in more detail in Chapter 5. Note that the two-level distinction also facilitates the theoretical formalisation of the comprehension and production processes and their relation to the grammatical knowledge of speakers. As part of the comprehension process, a hearer-reader must construct the hierarchical syntactic structure of the sentence given the words and their linear order. Given a syntactic structure, presumably elaborated from a coarse predicate-argument structure of an intended message, a speaker must linearise it into a string of words to produce an utterance. In this simplified description, word order acts as input for comprehension and as output for production processes while the hierarchical structure is the output for comprehension and the input for production.

We study phenomena of word order from the perspective of production, as observed in corpus data. Following the theoretical underpinnings presented above, our work adopts a simplifying assumption that word order is generated from and conditioned on the underlying dominance structure given by an unordered dependency tree.

Crucially, the mapping between the syntactic structure and its linearisation is not trivial and is language dependent. While in many constructions grammatical constraints define unambiguously the linear order of constituents (2.2a-b), we are interested in those cases and languages where there exist several possible grammatical orders (2.2c-d).

(2.2) a. John$_{sbj}$ likes Mary$_{obj}$

b. *John$_{obj}$ likes Mary$_{sbj}$

c. Джон$_{sbj}$ любит Марию$_{obj}$
John-NOM loves Mary-ACC

d. Марию$_{obj}$ любит Джон$_{sbj}$
Mary-ACC loves John-NOM

We only consider cases of word order variation where the linearisations express exactly the same syntactic structure (e.g., *John* is subject and *Mary* is object in example (2.2)).

The work presented in this thesis follows an empirical, corpus-based approach to word order variation which we present in the next section. We study constructions observed in corpora through the distribution of their occurrences. We focus on a class of factors affecting word order variation which are conditioned on the syntactic structure, that is, lengths of dependencies. By contrast, we do not address in detail the factors in word order variation at other linguistic levels (semantics, discourse, phonology).

2.2.1 Corpus-based empirical approach to word order variation

The large-scale development of syntactically-annotated corpora has inspired a resurgent interest for empirical syntactic work. The research methodology for word order variation analyses consists in extracting the cases of alternation of a chosen construction, annotating them with features corresponding to potentially relevant factors, and conducting a statistical analysis of the constructed dataset to test the effects of the hypothesised factors. The extraction of a set of constructions is typically semi-automatic: first, a candidate set is extracted using a pattern-matching of the syntactic annotations of the sentences. Afterwards, this dataset is verified and cleaned manually. The features which could not be obtained from the syntactic annotation (phonological, semantic, discourse information) are often added manually to the constructed corpus. Manual cleaning and augmentation of the extracted corpus are

typically possible on relatively small datasets containing between several hundred and a couple of thousands of observations.

The work of Gries (2001, 2003) is one of the representative examples of this approach. It is one of the first papers to propose that the hypothesised factors in a variation should be analysed jointly in a multifactorial statistical model instead of testing the effect of the factors separately, one by one.

Gries analyses the case of verb-particle split construction in English (2.3). Two orders are possible: with the particle adjacent to the verb (2.3a) and with the object intervening between the verb and the particle (2.3b).

(2.3) a. John *picked up* [the book]
b. John *picked* [the book] *up*

He uses the British National corpus and extracts 403 examples of this construction. As Gries discusses in his paper, the previous studies coming from different linguistic traditions and fields provide many analyses of the verb-particle variation. The variables that have been proposed include syntactic properties of the construction: the type of the direct object noun phrase (pronoun, definite, indefinite), the length and syntactic complexity of the direct object; semantic properties: e.g., how idiomatic the verb phrase is; phonological properties: where the stress falls on the verb, the length in syllables of the noun phrase; discourse properties: e.g., the information status of the direct object (given or new). Overall, Gries identifies and includes in his analysis 20 variables previously identified in the literature to explain patterns in the verb-particle variation. The values of the variables were annotated for each example sentence manually.

The resulting dataset consisting of 403 data points with 20 variables is analysed using several statistical tools: a classifier, a linear discriminant analysis and a logistic model. The use of logistic models is wide-spread in statistical analyses of word order variation since there are typically two competing word orders and the predicted variable is, therefore, a binary variable. In Chapter 4, we discuss more in detail generalised linear models and their extensions which we use for our analyses. For

each variable in the verb-particle variation, logistic regression gives a coefficient which determines its effect on the choice between the V Prt NP (2.3a) and V NP Prt (2.3b) orders. Gries found that the length of the object noun phrase has the strongest independent effect on the variation, with longer objects driving the preference for the V Prt NP order.

As can be seen from the work of Gries (2001, 2003) and other methodologically-similar studies (Arnold et al., 2000; Bresnan et al., 2007), the main challenges for exhaustive corpus-based analyses of syntactic variation lie in the manual cleaning and annotation of the data with the information on many relevant factors. In our work, we opt for fully automated analyses which focus only on a subset of factors. Among the factors which can affect word order variation, we analyse only the ones which can be extracted automatically from the syntactic annotation of the sentences. The advantage of this approach is that we can study variation at much larger scale: in larger samples of observations and across a number of languages.

Bresnan et al. (2007) is an important and influential work, providing a number of arguments in favour of adopting corpus-based statistical analyses as tools and data for theoretical syntax. Bresnan et al. (2007) investigate the case of dative alternation in English (2.4).

(2.4) a. ... gave [toys] [to the children]
b. ... gave [the children] [toys]

The alternation in this construction is between the order of the theme (*toys*) and the recipient (*the children*) with the recipient-theme order resulting in double object structure (2.4b). Bresnan et al. (2007) analyse many factors affecting this variation, including the relative length of the theme and the recipient, the animacy of the recipient, the given-new status of the theme, and others. They find that most of the factors have an independent effect on the choice of order in dative alternation. Similarly to the results reported by Gries (2001), the relative length of the phrases is one of the prominent factors identified: a longer theme phrase tends to be placed after a shorter recipient and vice versa. Importantly, Bresnan et al. (2007) show

that the effects they find hold in two different corpus samples: spoken spontaneous speech and written journalistic texts. This result is taken to indicate that the factors affecting dative alternation are part of the general probabilistic syntactic knowledge of speakers and not simply a contingency of a particular corpus, genre or modality.

One of the experiments carried out in Bresnan et al. (2007) concerns the lexical biases in dative alternation. Semantic properties of the theme and recipients arguments often depend on the verb semantics. For example, the recipients of the verb *bring* express a given referent much more often than the recipients of the verb *take*. To account for this lexically-conditioned variation, Bresnan et al. (2007) employ a novel statistical technique for word order variation analysis: a logistic mixed effect model (Pinheiro and Bates, 2000; Bates et al., 2014), an extension to the logistic and linear models. Using this model, the authors show that the effects of the factors observed by fitting a logistic regression are independent of lexical properties of verbs. We adopt the statistical model and methodology from Bresnan et al. (2007) for our experiments reported in Chapter 4 to take into account the lexically-conditioned part of the adjective variation in Romance languages. We provide a detailed description of generalised mixed effect models in Section 4.1.2.

2.3 Dependency length minimisation

Dependency length minimisation (DLM) is a term that we use in this work to refer to a number of related tendencies in language production as observed in corpus data, psycholinguistic experiments and typological variation. This section provides an overview of research on DLM with a focus on corpus-based studies.

"Languages tend to put related words in a sentence close to each other" is one possible way to state DLM very generally at the grammatical and typological level of description. This tendency has been noted in such form already by Behaghel (1932). In one form or another, adjacency principles have been used to account in a systematic way for many word order patterns in languages. The fact that a prepositional phrase cannot intervene between a verb and its object complement (**John ate with a fork a cake*) can be explained, for example, in terms of a constraint on the adjacency of syntactically and

semantically dependent elements (verb and object). In the typological variation, an adjacency principle was used, for example, to account for the order of modifiers in the noun phrase (Rijkhoff, 1998). Based on the assumption that adjectives should be more semantically related to the noun than numerals, Rijkhoff (1998) concludes that there should be a preference toward word orders such as Num Adj N, Adj N Num or N Adj Num where the noun and the adjective are adjacent, and against orders such as Adj Num N or N Num Adj, where the adjective and the noun are placed apart. While these predictions are born out to some extent in a typological sample of languages, the whole picture was shown to be much more complicated (Dryer, 1992; Cinque, 2005).

The most relevant for our work is the other type of evidence for DLM effects which comes from word order variation studies. We review in the following section the well-documented phenomena such as "heavy-NP shift", "short-before-long tendency" and other, gathered under the umbrella of the DLM principle.

2.3.1 DLM effects in word order variation

The most frequently studied constructions in English in connection with the minimisation of lengths of dependencies are "rightward" shifts in the canonical placement of heavy noun phrases and other related alternations in postverbal domain (Wasow, 2002). These constructions typically involve two dependents following the head verb and include verb-particle split ((2.5), Gries (2001); Lohse et al. (2004)), dative alternation ((2.6), Bresnan et al. (2007)), heavy-NP shift ((2.7), Ross (1967); Stallings et al. (1998)), ordering of multiple prepositional phrases ((2.8), Hawkins (1999); Wiechmann and Lohmann (2013)).

(2.5) a. throw the trash out
b. throw out the trash

(2.6) a. give a book to Mary
b. give Mary a book

(2.7) a. reveal the news at dawn

b. reveal at dawn the news

(2.8) a. arrive at work on Thursday

b. arrive on Thursday at work

The first order is typically the preferred, canonical one in these constructions, but the preferences can change if some of the constituents become longer, or heavier, phrases.

(2.9) a. throw [the bin with old trash] out

b. throw out [the bin with old trash]

(2.10) a. give [a book which I have bought] to Mary

b. give Mary [a book which I have bought]

(2.11) a. reveal [the news about the merger with a competitor] at dawn

b. reveal at dawn [the news about the merger with a competitor]

(2.12) a. arrive [at the new office branch] on Thursday

b. arrive on Thursday [at the new office branch]

The preference for the second order in the examples (2.9–2.12) can be described as a heavy phrase shift (Stallings et al., 1998), since a phrase (e.g. the noun phrase *a book which I have bought* in (2.10)) moves rightwards from its canonical, immediately postverbal, position in the sentence. Other related constructions in English that can be seen as a heavy phrase shift include extrapositions such as *a woman appeared* [*who was not invited*].

However, a more accurate generalisation of these data can be stated as a "short-before-long" preference, that is: "shorter phrases tend to occur before longer phrases". This generalisation takes into account the fact that often both verb dependents (2.10–2.12) can be long phrases and it is the difference in their lengths which influences the preferred order. Evidence for the relative length generalisation is given by corpus investigations (Hawkins, 1994; Wasow, 1997; Hawkins, 1999; Bresnan et al., 2007) which found that the corpus frequency of the option *b* in the examples (2.9–2.12) is correlated with the length of the shifted phrase relative to the length of the

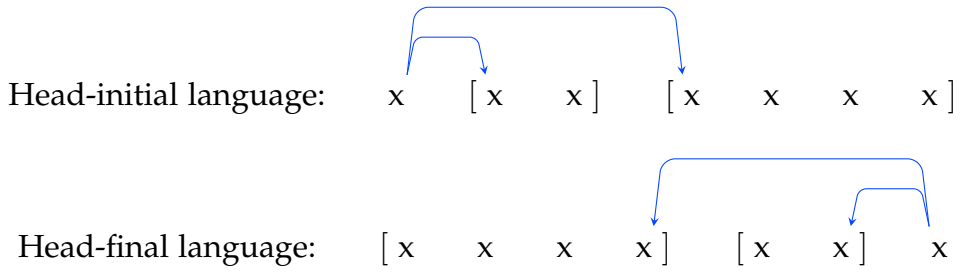


Figure 2.7: Illustration of "short-before-long" principle for head-initial and "long-before-short" principle for head-final languages.

second dependent. Additional evidence comes from experimental data in the form of acceptability judgments (Wasow and Arnold, 2003) and production frequencies (Stallings and MacDonald, 2011).

While this generalisation is quite accurate for English, it does not hold in the same way in head-final languages such as Japanese. In fact, Japanese data show the preference for "long-before-short" ordering of phrases in the verbal domain (Hawkins, 1994; Yamashita and Chang, 2001). Since Japanese is a head-final language, the verb in the sentence is placed at the end, after its dependents. Similarly, noun and prepositional phrases have their heads at the right edge of the phrase. Thus, Japanese verb phrase shows the exact mirror ordering to that of the English verb phrase (Figure 2.7). Taking together the evidence from head-initial languages and head-final languages (preferences similar to those in Japanese were found in Korean (Choi, 2007) and Basque (Ros et al., 2015)) we can reformulate the previously proposed heavy-XP shift or "short-before-long" principle in terms of a more general distance or dependency length minimisation effect. As can be intuitively seen from the illustration in Figure 2.7, if we take as relevant properties the distances between the phrases and their head (the verb), the generalisation can be stated cross-linguistically as the minimisation of the distances between the head and its dependents, or, equivalently, the lengths of the dependencies.[12] The work of Hawkins (1994, 2004) provides substantial evidence for the generalisation of these word order alternation data based on dependency length. It is also one of the first worked-out explanations of the DLM effects in terms

[12]Section 4.2 further discusses the relation between "short-before-long" principle and the dependency length minimisation formulation.

of processing bias.

2.3.2 Processing accounts of DLM effects

The processing account outlined in Hawkins (1994) provides a metric to calculate online parsing complexity based on an intuitive idea that processing distant dependent constituents should be harder than processing constituents close to the head. The explanation for parsing complexity lies, in its turn, in the direct correlation with working memory limitations of the human sentence processing mechanism.

The metric of processing complexity is based on the idea that processing a phrase and constructing its parsing structure requires identification of its immediate constituents (ICs) — the head node and the syntactic types of its daughter nodes — and includes processing of all constituents between the first and the last IC (Constituent Recognition Domain, CRD, (Hawkins, 1994, 58)). For example, the parser can construct the verb phrase (VP) in the heavy-NP shift example (2.13a) only after identifying the head verb *gave*, the daughter noun phrase (after seeing the determiner *the*) and the prepositional phrase (after seeing the preposition *to*):

I [$_{VP}$ gave [$_{NP}$ the valuable book that was difficult to find] [$_{PP}$ to Mary]] (2.13a)

1 2 3 4 5 6 7 8 9 10

I [$_{VP}$ gave [$_{PP}$ to Mary] [$_{NP}$ the valuable book that was difficult to find]] (2.13b)

1 2 3 4

As the two variants of the same sentence illustrate, depending on the ordering of constituents, the parser might be able to build the VP structure very early during the processing (2.13b) or relatively late (2.13a). According to the principle of Early Immediate Constituents (EIC, (Hawkins, 1994, 77)), "the human parser prefers linear orders that maximise the IC-to-non-IC ratios of constituent recognition domains". In other words, it is best for the human parser to minimise the average processing time

(that is, the number of observed words) necessary to identify the structure of a phrase. For the example in (2.13a and 2.13b), the principle of EIC predicts that the sentence in (2.13a) will be harder to process than the sentence in (2.13b) as the constituent recognition domain of the VP in (2.13a) includes ten words, while it contains only four words in (2.13b).

EIC computation works similarly for head-final languages (Hawkins, 1994, 80). The number of words to process is counted from the first phrase head, e.g., the determiner *the*, to the last phrase head, i.e., the verb. Since the heads of the dependent phrases are phrase-final, the computation of IC-to-non-IC ratio proceeds in a way which is the exact mirror to the head-initial scenario (Figure 2.7).

The Minimisation of Domains principle (Hawkins, 2001, 2004) is an extension of the EIC principle which proposes that both syntactic and semantic dependencies are susceptible to minimisation. As such, this principle is approaching a very general formulation of the DLM principle which we cited at the beginning of this section (Behaghel, 1932). For instance, (Hawkins, 2004) uses semantic and syntactic dependency minimisation to explain why complements are more adjacent to heads than adjuncts. In reality, he restates the syntactic constituency principles from a processing perspective.

While the processing principles proposed by Hawkins are initially developed for parsing, he attempts to propose that these principles are more general and apply in a similar way to language production. This assertion is motivated only empirically, that is, by the fact that the DLM effects are observed in word order variation, including corpus data and experiments with elicited production. The overall picture painted by Hawkins suggests that the DLM principle is a manifestation of the properties of one processing mechanism operating both in production and comprehension.

Gibson's Dependency Locality Theory (DLT, Gibson (1998, 2000)) suggests a related processing account of DLM. DLT is developed as a general theory of syntactic complexity to account for phenomena in sentence processing such as difficulties in comprehension of nesting clauses and the relative difficulty of processing object- versus subject-extraction. Similarly to Hawkins' work, DLT provides a measure of

complexity of a sentence based on its syntactic structure. The syntactic complexity of a sentence is a sum of two components: storage cost and integration cost. Storage cost computes the memory requirements for maintaining the partially-built syntactic structure and the requirements that should be yet satisfied. Integration cost computes the processing load of activating previous words which should be connected to the currect word (i.e., assigning a head to a dependent or vice versa). Importantly, integration cost increases with the distance between the previous and the current words being integrated. Intuitively, this is because the activation of the stored elements decreases over time. The (re)activation is harder the longer the stored element was inactive. More precisely, the distance and activation decay is measured in terms of a number of new discourse elements introduced between the last activation of the stored lexical element and the current word.

As Gibson (1998, Section 4) discusses, distance-based integration cost explains the preference for short-before-long order in word order alternations such as heavy-NP shift in English. Gibson explains the preference between two possible word orders in terms of judgments of intuitive complexity and suggests that the relevant measure is the maximal integration cost (as opposed to, e.g., the average cost). Hawkins' and Gibson's proposals differ in conceptual explanation (minimisation of the time to construct a constituent versus minimisation of activation time) and the definitions of distance between the dependent elements (lengths in number of words versus lengths in number of discourse references). Despite this, the overall argument and the computations involved in deriving the short-before-long (or long-before-short for head-final languages) principle are very much similar in DLT to the ones applied in EIC principle. Compare the computations in examples (2.13a) and (2.14). The two lines in example (2.14) indicate the units of storage cost and integration cost. We are interested in integration cost which differs for the two possible orders. The maximal integration cost is 3 in the example (2.14) because at the moment of attaching *Mary* to the verb *gave* there are three discourse referents (two verbs and one noun) which were introduced in-between, decaying the activation of the verb. A similar computation for the order (2.13b) results in maximal integration cost of only 1, suggesting that the order (2.13b) has lower processing complexity and should be preferred against the

order (2.13a), exactly as in Hawkins' account.

I [$_{VP}$ gave [$_{NP}$ the valuable book that was difficult to find] [$_{PP}$ to Mary]] (2.14)

1	0	0	1	0	1	0	0	1	0	1
0	0	0	1	0	0	0	0	0	0	3

2.3.3 Recent large-scale work on DLM

DLT is primarily concerned with explaining comprehension data. In important follow-up work, Temperley (2007) tests systematically the predictions of DLT in production using corpus frequencies. The objective of this work is to apply DLT computations to varied constructions in English and test its predictions statistically in a treebank. To do so, Temperley (2007) simplifies and generalises DLT in several important ways, propelling the use of a more general "dependency length minimisation" principle. First, he measures the distance between dependent elements simply as a number of intervening words (as opposed to a number of discourse referents). Secondly, the complexity of a sentence or a phrase is measured as a total length of all its dependencies. This definition corresponds to associating complexity with the (total) processing time of a sentence as opposed to the peak value of memory or computational load encountered during processing. One piece of evidence collected in favour of this interpretation of complexity concerns the ordering of more than two postverbal elements. If maximal processing load (or, in Hawkins' EIC principle, the overall constituency domain) is correlated with complexity, then only the dependency between the head and the most distant dependent should be minimised. The order of the other two dependents does not need to be optimised and follow the short-before-long principle. Contrary to this prediction, Temperley (2007) shows that the first dependent tends to be shorter than the second dependent even when a third dependent is present.

The formulation of a DLT-inspired complexity criterion for preferences in production allowed Temperley to make quantifiable predictions for several syntactic variation constructions, expanding substantially on the traditional data of alternations in the

postverbal domain. For instance, one prediction concerns the lengths of subject versus object noun phrases. Since noun phrases are mostly right-branching, the distance between the verb and the subject is approximately the length of the subject phrase. On the other hand, the distance between the verb and its object does not depend on the object length. Consequently, dependency length minimisation predicts that subject noun phrases should be on average shorter than object noun phrases, which is confirmed statistically. Similarly, adverbial clauses appearing before the subject and the verb tend to be shorter than adverbial clauses appearing after the verb. The general DLM prediction confirmed by these data can be stated as "In a primarily right-branching language, the left-branching constituents should be short." (Temperley, 2007, p. 306). Interestingly, Temperley also reports some cases which do not support the DLM principle. In particular, the predicted long-before-short ordering of pre-modifying adjuncts is not confirmed by the corpus data. Instead, a slight short-before-long tendency is observed, suggesting the presence of other factors which interact with dependency length constraints and affect word order choices.

The generalised approach proposed by Temperley (2007) to test DLM predictions in syntactic choice constructions based on the lengths of dependencies in a treebank was adopted and extended in several subsequent papers (Gildea and Temperley, 2007; Temperley, 2008; Park and Levy, 2009; Gildea and Temperley, 2010). These studies take one step further to quantify the presence of DLM in a language overall, i.e., by comparing the dependency lengths of English sentences with dependency lengths in parallel sentences with a permuted order of words. The idea is that, given a dominance structure (provided by the treebank annotation), there exist an order of words which minimises the total length of dependencies in it. By comparing the average length of dependencies between the original word order, the "optimal" word order, and a random, not optimised, word order, one can provide quantitative evidence for DLM on "average", for the language as a whole. This approach treats all dependencies equally, and the word order of each sentence is optimised independently.[13]

Using such quantitative comparisons, Gildea and Temperley (2010) found that English

[13] A comparison with the "optimal" linearisations preserving the direction of dependency across sentences was also proposed (Gildea and Temperley, 2007).

minimises dependencies to a large degree, with average dependency length of 2.24 being relatively close to the optimal dependency length of 1.58 (WSJ corpus). On the other hand, the German treebank used in the analysis has the average dependency length of 2.95 further away from the optimal dependency length of 1.56.[14] Futrell et al. (2015b) conducted similar experiments on treebanks of 40 languages and found that all languages minimise dependencies compared to average dependency lengths of random order permutations but the extent of this minimisation is variable. For instance, perhaps unsurprisingly, languages with free word order such as Latin and Ancient Greek have relatively long on average dependencies.

In another type of large-scale analysis of DLM effects, Rajkumar et al. (2016) re-examine the preferences in variation constructions studied by Temperley (2007). Methodologically, Rajkumar et al. (2016) follow the path of traditional statistical approaches to word order variation described above (Gries, 2003; Bresnan et al., 2007) but extend them in two interesting ways. First, the authors simultaneously examine the variation of four different types of constructions in relation to the same set of factors. Secondly, they use one single model for the four constructions which is optimised jointly on the combined dataset of patterns. To do so, instead of a simple logistic regression distinguishing between two alternative orders (e.g., 0 and 1), Rajkumar et al. (2016) use a logistic-based ranking model. The ranking approach allows one to define the choice between word order options without specifying distinct labels for each construction. For each word order observed in the corpus, potential alternative orders are constructed, and the model learns to rank the observed order higher than the alternatives, based on their sentence-level features.

Rajkumar et al. (2016) extend the work of Temperley (2007) by controlling for frequency-based factors in addition to dependency length effects. They include, in particular, the average n-gram log-probability of the words in a sentence and the log-likelihood of a sentence based on its probabilistic context-free grammar (PCFG) parse. These measures are inspired by expectation-based accounts of comprehension difficulty (Hale, 2001; Levy, 2008). In production, however, they mostly play a role of controls for cases when *a* is preferred over *b* because *a* is, in general, more frequent than *b*, regardless of the dependency lengths of the *ab* and *ba* orders. Similarly to

[14] See also Park and Levy (2009) for a similar finding.

Temperley (2007), Rajkumar et al. (2016) use the total dependency length of a sentence as an independent variable in their analysis. They investigate several variants of dependency length: computed as a number of discourse referents (as in Gibson's DLT), as a number of words (as in Temperley's interpretation of DLT) and as a number of stressed syllables. They find that the three measures produce very similar accuracies in predicting the correct word order option on the Brown and Wall Street Journal (WSJ) corpora. The differences are not significant for the Brown corpus, but the discourse referent measure is slightly but significantly more accurate on the WSJ corpus. Consequently, they adopt Gibson's measure of dependency length for their analyses.

The primary outcome of the study of Rajkumar et al. (2016) is the confirmation of the findings of Temperley (2007) controlling for frequency effects. Thus, they provide robust evidence for the dependency length minimisation in several constructions of word order variation in English. The effect of dependency length and other factors are, interestingly, very similar in the two corpora that were examined. This result suggests that genre does not affect syntactic properties of word order variation, supporting a similar observation by Bresnan et al. (2007). As in Temperley (2007), Rajkumar et al. (2016) also find that not all constructions follow DLM. The ordering of preverbal adjuncts leans towards the short-before-long order instead of the long-before-short order predicted by DLM.

2.3.4 Summary

The previous work reviewed in this section provides substantial evidence for DLM effects in word order variation phenomena. However, the data examined remain often limited to English and a few constructions of similar type, e.g., alternations in the postverbal domain. The recent work based on cross-linguistic treebank data put forward an approach to analyse DLM effects on a large-scale, across many types of dependencies and constructions. This thesis contributes to the growing body of work in this direction. We analyse and refine the approach of Gildea and Temperley (2010) to measure average dependency lengths of sentences (Chapter 3). We extend the data

relevant for DLM theories by analysing the adjective variation in Romance languages using the treebank-based general DLM formalisation (Chapter 4).

Chapter 3

The DLM principle and word order variability at the language level

This thesis presents computational analyses of word order variation and dependency length minimisation (DLM). The presentation will proceed in a top-down fashion. In this chapter, we start by looking at these two aspects of word order at the top level of representation, that is, as observed in a language overall.

The language-level quantitative approach to studying word order properties is a new direction in computational linguistic research, inspired by the emergence of multilingual syntactic treebanks. This approach is useful from two perspectives. First, the parameterisation of word order is instrumental for classification and comparison of languages in typological studies. Second, typological and quantitative properties of languages inform the development and adaptation of NLP systems, such as statistical parsers, for multilingual tasks. Some fundamental questions that the quantitative approach can help answer are "Which word order properties are important for defining a cross-linguistic typology?" and "Which languages have freer word order than others?" In the context of NLP applications, it is crucial to know which word order properties are relevant for a particular task, how word order properties affect the performance of a system and for which languages a system should be expected to obtain similar results.

Our experiments expand on the previous work in this emerging field in two ways.

First, we show that we can quantify dependency lengths in a treebank to compare how much languages minimise their dependencies (Section 3.1). Our case study of several texts in Latin and Ancient Greek serves as an illustration of the proposed DLM ratio measure and its properties and limitations. The treebanks come from the PROIEL project (Haug and Jøhndal, 2008) and contain four texts in Latin (from the 1st century BC and the 4th century AD) and two in Ancient Greek (from the 4th century AD and 4th century BC). We work with these ancient languages because they have relatively free word order and long dependencies providing us with enough variability for interesting empirical analysis. In addition to dependency length, we analyse and quantify other word order properties in Latin and Ancient Greek texts. We focus, in particular, on word order variability computed as arc-direction entropy.

Secondly, we demonstrate that the DLM and word order variability measures are useful in practical NLP applications, specifically, as word order correlates of parsing performance (Section 3.2). Statistical dependency parsers are widely assumed to perform worse on longer dependencies and languages with relatively free word order. The treebanks of Latin and Ancient Greek provide us with an opportunity to test this claim in a controlled setting. To further extend the analysis of the effect of word order properties, quantified as DLM ratio and arc-direction entropy measures, on parsing performance, we propose to evaluate artificial treebanks constructed by permuting sentences of natural language treebanks.

3.1 Measuring DLM and word order variability in a treebank

In this section, we present and analyse several measures of dependency lengths and word order variability in a dependency treebank. Our exposition is based on the discussion of the previous work which has proposed several related but slightly different measures. Our choice of the measures is motivated by the previous findings and the theoretical and empirical considerations about their statistical properties. We

illustrate the empirical properties on the PROIEL treebanks of Latin and Ancient Greek which we introduce below.

3.1.1 Corpus data for empirical analysis: Latin and Ancient Greek PROIEL treebanks

The empirical results we report in this section, concerning measurements of dependency length and word order variability, are obtained for Latin and Ancient Greek languages based on their treebanks. We chose these languages for our case studies for several reasons. First, both Latin and Ancient Greek allow much freedom in the linearisation of sentence elements. Secondly, they are traditionally extensively documented and curated and have been syntactically annotated more recently (Bamman and Crane, 2008; Haug and Jøhndal, 2008). Compared to modern languages that have treebank resources, these two ancient languages have greater word order variability. While there is an increasing treebank coverage of the languages of the world, the majority of the existing treebanks have been developed for Indo-European languages and widely spoken languages such as Chinese, Arabic and Japanese. Unfortunately, the indigenous languages featuring interesting word order properties do not yet have necessary treebank resources for quantitative analysis.

Also, we have access to several texts in the Ancient Greek and Latin treebanks that were written in the same language but that have different word order properties, since they come from different historical periods. The comparison of word order between these texts provides some interesting observations about word order properties and diachronic change.

The dependency treebanks of Latin and Ancient Greek used in our study come from the PROIEL collection (Haug and Jøhndal, 2008). The PROIEL corpora contain exclusively prose and is, therefore, more appropriate for a word order variation study than previously developed treebanks, such as the Perseus treebanks (Bamman and Crane, 2011), which also contain poetry. Moreover, the PROIEL collection allows us to analyse different texts and authors independently of each other. Table 3.1 presents the

Language	Text	Period	#Sent	#Words
Latin	*Caesar*, Commentarii belli Gallici	58–49 BC	1154	22408
	Cicero, Epistulae ad Atticum & De officii	68–43 BC	3830	44370
	Peregrinatio Aetheriae	4th c. AD	921	17554
	Jerome's *Vulgate*	4th c. AD	8903	79389
Ancient Greek	*Herodotus*, Histories	450–420 BC	5098	75032
	New Testament	4th c. AD	10627	119371

Table 3.1: Summary of the properties of six Latin and Ancient Greek treebanks, including the historical period and size of each text. Italic indicates the short names we will be using for the texts.

texts included in the corpus with their time periods and their number of sentences and words.

The two Greek texts are Herodotus' Histories (5th century BC) and the New Testament (4th century AD). Two of the texts in Latin are from the Classical Latin period (Caesar and Cicero), and the other two are in the Late Latin of the 4th century (Vulgate and Peregrinatio). Note that Jerome's Vulgate is a translation of the Greek New Testament. The sizes of the texts are uneven, but each includes at least 17'000 words or 900 sentences.

The dependency annotation scheme is similar to the Universal Dependencies treebanks, as we mentioned in Chapter 2. We implemented an automatic procedure to extract the lengths of dependencies and other statistics on dependency arcs. This procedure is straightforward given the syntactic information contained in the annotation, such as the word and head indices.

3.1.2 Measuring the degree of DLM

We introduced the DLM principle as a general way to refer to tendencies observed in language production to place syntactically related words and constituents close to each other in the linear order of an utterance. We distinguish further two groups of minimisation effects: those we observe at the level of grammar and those we find in

word order alternation preferences. Typological data on word order constraints across the languages of the world has long provided evidence that frequently found grammar types produce shorter dependencies than rarely found grammars (Greenberg, 1963; Hawkins, 1983). The corpus data have been the frequent source of evidence for DLM effects at the level of word order variation.

Recently, the corpus data has been leveraged as evidence for DLM at the language level, without explicitly distinguishing between the effects in the grammar or in the variation (Temperley, 2008; Liu, 2008; Gildea and Temperley, 2010; Futrell et al., 2015b). These studies make use of dependency treebanks that make it possible to compute dependency length automatically for a large number of sentences and all types of dependencies. While a dependency length statistics at the treebank level confounds several distinct DLM effects, it serves as a systematic way to analyse the general tendency of languages to minimise dependencies. For instance, using dependency treebanks from 40 languages, Futrell et al. (2015b) concluded that there is an overall tendency toward DLM cross-linguistically. However, languages minimise dependencies to varying extents. In this work, we take this one step further and argue that the quantification of dependency length at the treebank level can be used to compare the degree of DLM across languages.

We use the dependency length $DL(s)$ of a sentence s as our basic measure, following the previous analyses of DLM at the treebank level.[1] Formally, take a sentence $s = w_1, \ldots, w_n$ annotated with its dependency tree structure t_s, where indices $i = 1, \ldots, n$ give the linear order of words. Then, the length of a dependency $d \in t_s$ between words w_i and w_j is equal to $|j - i|$. For instance, the length of a dependency between adjacent words is equal to 1. The dependency length of a sentence is the sum of the lengths of all its dependencies:

$$DL(s) = \sum_{w_i \to w_i \in t_s} |j - i| \tag{3.1}$$

In the sentence in Figure 3.1, there are five dependencies whose individual lengths — computed as the difference between word indices — are indicated above the

[1] Alternatively, one could use the average length of a single dependency $\langle DL(d) \rangle$ as a basic measure. Note that $DL(s) = (n-1) \cdot \langle DL(d) \rangle$ where n is the length of the sentence.

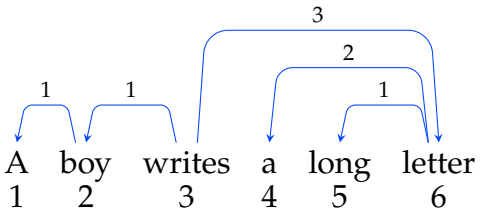


Figure 3.1: Illustration of the computation of dependency lengths. The DL of this sentence is 8.

dependency arcs. The total dependency length of the sentence is 8.

The DL of a sentence depends on the sentence length n and the tree structure t. Longer sentences have more dependencies (equal to $n - 1$) and, consequently, a higher total DL. If we want to compare languages with respect to their dependency length at the treebank level, we cannot compute a simple average of DLs across all sentences in a treebank as was done, for example, in the work of Gildea and Temperley (2010). Such a measure will depend on the composition of a treebank by sentences of different lengths creating a confounding factor in comparisons across treebanks.

Even more crucially, two sentences with the same length can have different DLs because of the particular syntactic structures they are generated from. For example, the two sentences in Figure 3.2 have slightly different tree structures: the head verb *writes* has two children in the first sentence and three children in the second sentence. The resulting dependency lengths of the sentences are also different: $DL(s_1) = 5$ and $DL(s_2) = 6$. Generally, a larger average number of children per head, known as the branching factor of a tree, leads to a larger DL of a sentence (Ferrer-i-Cancho, 2013).

This example illustrates a conceptual problem: an absolute dependency length measure is not appropriate to compare whether one sentence (or treebank or language) minimises dependencies more than some other sentence. This problem has been recognised in previous work that proposed to compare the DLs of sentences with their minimal possible DLs (Gildea and Temperley, 2007; Park and Levy, 2009; Gildea and Temperley, 2010; Tily, 2010; Futrell et al., 2015b). Gildea and Temperley (2007,

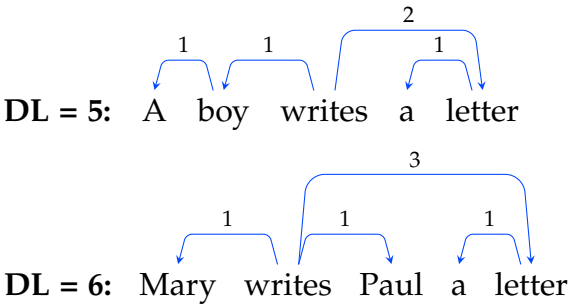


Figure 3.2: Illustration of difference in DLs for two sentences with the same length.

2010) proposed an algorithm to produce, given an underlying syntactic structure of a sentence, a linear order of the words that yields the minimal possible DL. As an illustration, consider the minimal DL of the two sentences from the previous example (Figure 3.3). The words in the first sentence can be placed so that the lengths of the dependencies sum up to 4 (DL of the original linearization is 5). Note that the dependencies (e.g., *boy* → *a*, *writes* → *boy*, *writes* → *letter*, etc.) remain exactly the same as in the source sentence (Figure 3.2). Similarly, an optimal — from the perspective of DLM — linearization for the second sentence yields a DL equal to 5, smaller than the DL of the original word order which was equal to 6. Importantly, there is a difference between the minimal DLs of the two sentences which reflects the fact that the underlying tree structure constraints possible linearizations and can lead to long dependencies in some cases but not in others.[2]

In addition, previous work has compared the DLs of observed linearizations to the DLs of random linearizations of words in a sentence (Ferrer-i-Cancho, 2004; Liu, 2008; Park and Levy, 2009; Gildea and Temperley, 2010; Futrell et al., 2015b). The optimal DL serves as a lower bound while the random DL serves as an upper bound on possible DL values given a tree structure. By placing the DLs of actual sentences between these upper and lower bounds, one can observe whether there is a DLM effect at the treebank level. Gildea and Temperley (2010) observed that actual DLs are

[2] An implicit assumption behind the comparison of an observed word order with an optimal order constructed as a permutation of words in the sentence is that DLM operates on a fully generated underlying syntactic representation. In other words, the language production mechanism generates a syntactic structure of a sentence which is then linearised through some principles including the DLM principle. This interpretation is in line with our initial assumptions on word order as a separate part of the grammar on top of unordered hierarchical representation.

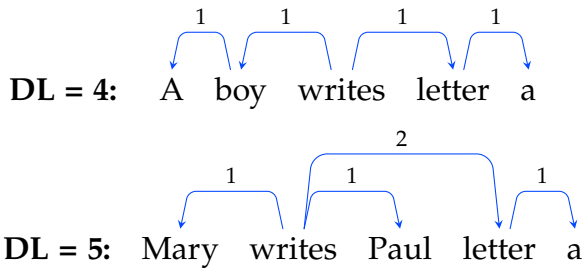


Figure 3.3: Illustration of difference in optimal DLs for two sentences with the same length.

very close to optimal DLs for English but less so for German, a language known for its long dependencies. On a sample of 40 languages, Futrell et al. (2015b) confirm that all languages tend to have shorter dependencies compared to their randomly permuted counterparts, but the extent of this minimisation varies across languages. For illustration, in Figure 3.4, we plot the DL, random DL (RandDL) and optimal DL (OptDL) measures for the six PROIEL treebanks, averaged across sentences of the same length. We can observe that the DL values of each are smaller than the random upper bound but also larger than the optimal lower bound. Futrell et al. (2015b) focus on the importance of cross-linguistic confirmation of the DLM principle as observed in Figure 3.4 from the difference between the curves. Instead, our goal is to estimate, using the same data, which treebanks in Figure 3.4 minimise dependencies more than the others.

Figure 3.4 also highlights the fact that DL measures depend on the sentence length. In particular, Ferrer-i-Cancho (2004) showed theoretically that the average random DL is distributed as a function of n^2. The relations of the actual and optimal DLs to sentence length cannot be established theoretically in a similar manner, since they depend, in particular, on the average branching factor of the tree (Ferrer-i-Cancho, 2013).[3] This means that the average dependency length across all sentences in a treebank cannot be directly compared to the average minimal and random DLs (e.g., as previously proposed by Gildea and Temperley (2010)). Ferrer-i-Cancho and Liu

[3] Based on the empirical distributions in the treebanks, Futrell et al. (2015b) suggest quadratic approximations for the three curves. However, our empirical tests and the theoretical considerations in (Ferrer-i-Cancho, 2013) point to a subquadratic and, possibly, linear relation between the optimal and actual DLs and the sentence length.

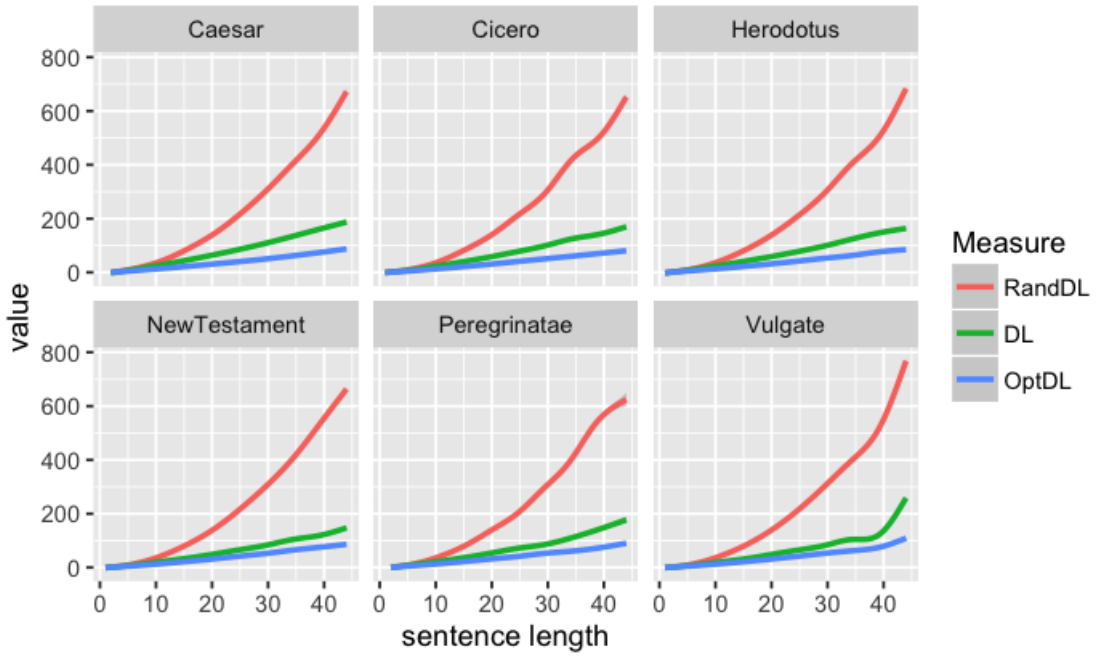


Figure 3.4: Actual (DL), optimal (OptDL) and random (RandDL) dependency length measures in PROIEL treebanks averaged across sentences of the same length.

(2014) make this point in a theoretical probabilistic analysis of dependency length distributions and their relation to the sentence length. Informally, as can be seen in Figure 3.4, the difference between the measures increases with sentence length.

We would like to suggest that a better measure for cross-linguistic quantification of DLM at the treebank level is the average ratio of the DL and the optimal DL computed for each sentence in a treebank:

$$DLMRatio = \frac{1}{k}\sum_{i=1}^{k}\frac{DL(s_i)}{OptDL(s_i)} \tag{3.2}$$

In other words, we propose to compute the *relative* average DL, normalised by the minimal possible DL.

Tily (2010, pp. 63–68) previously used this measure to track the rate of DLM in historical corpora of English. Despite the high variance in the DLM ratio across texts

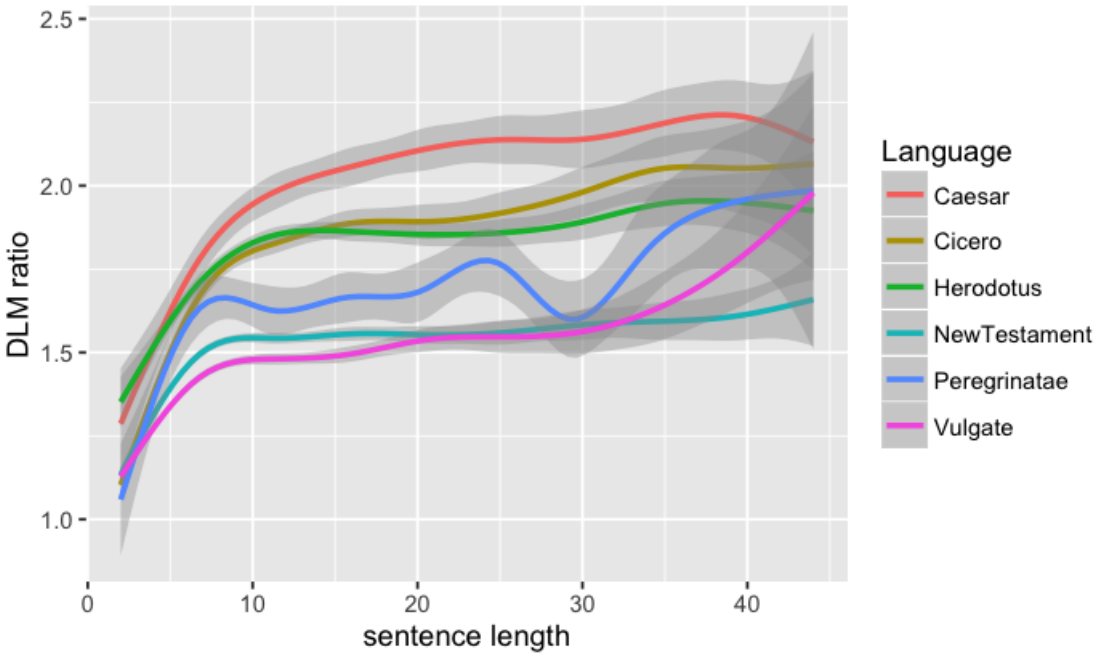


Figure 3.5: Average DLM ratio measure for sentences of different lengths in the PROIEL treebanks.

of the same period, he found a trend towards shorter dependencies from older to more recent English. Tily (2010) uses the ratio measure to confirm at a conceptual level the effect of DLM on language change; he does not analyse statistical properties of the measure nor its relation to the sentence length. Ferrer-i-Cancho (2004) computed the ratio measure for sentences in one treebank (Romanian) and found a slight tendency for the ratio to increase with the sentence length.

To demonstrate the validity of the DLM ratio measure for empirical cross-linguistic comparison of DLM effects, we analyse it on a collection of historical treebanks in Latin and Ancient Greek from the PROIEL collection (Haug and Jøhndal, 2008).

Empirical analysis of the DLM ratio in the PROIEL treebanks

First, we analyse the relation of the DLM ratio measure with the sentence length. Figure 3.5 plots the DLM ratio measure averaged for sentences of the same length in our six PROIEL treebanks. The OptDL factor in the DLM ratio is computed using

the algorithm of Gildea and Temperley (2010) summarised in Section 3.2.3. The grey regions around the fitted smoothed curves show standard errors in the estimation of the average DLM ratio.[4] As we can see from the plot, the DLM ratio depends (non-linearly) on the sentence length and does not immediately alleviate our basic problem of averaging across sentences of different lengths. Simplifying, we can describe the relationship as having two regimes: for short sentences of length 2–10, there is a clear increase in the DLM ratio, while for sentences of length greater than 10 we can consider the DLM ratio to be almost constant with sentence length. It is not, in fact, surprising that very short sentences tend to have small DLM ratios close to 1. For sentences of length 2, the only possible orders are both optimal with respect to dependency length. Similarly, for sentences of length 3, 4, 5 there should be a relatively high probability that an order is optimal given that there are not many possible orders. This can also be concluded informally from the fact that the average DL of random linearizations is very close to the optimal DL for very short sentences (see Figure 3.4).

Figure 3.6 demonstrates the histogram distribution of DLM ratio values in our treebanks. We can see a peak for the DLM ratio of value 1 in all treebanks, particularly noticeable in Cicero and Vulgate. Apart from this irregularity on the left edge of the distribution, the DLM ratio distribution is close to log-normal. As can be seen from the density diagram in Figure 3.7, if we eliminate all sentences shorter than 10 words, the remaining DLM values are approximately distributed log-normally.[5]

Overall, these empirical data suggest that we can measure the DLM ratio robustly starting with sentences of length 10. While very short sentences are harder to compare across languages and treebanks, we will assume in the following that the DLM ratio for longer sentences approximates well enough the total degree of DLM in a treebank, including both short and long sentences. Future empirical and theoretical analyses of the DLM ratio could clarify how the DLM principle affects DLs in short sentences.

Table 3.2 gives the treebank-level DLM measure — the average DLM ratio for sen-

[4] This and all other plots in this thesis are produced using R package `ggplot2` developed by Hadley Wickham.

[5] The sentence length threshold of 10 is based on the Figure 3.5. The distribution of DLM measure was very similar in our data starting from sentences of length 6 and longer.

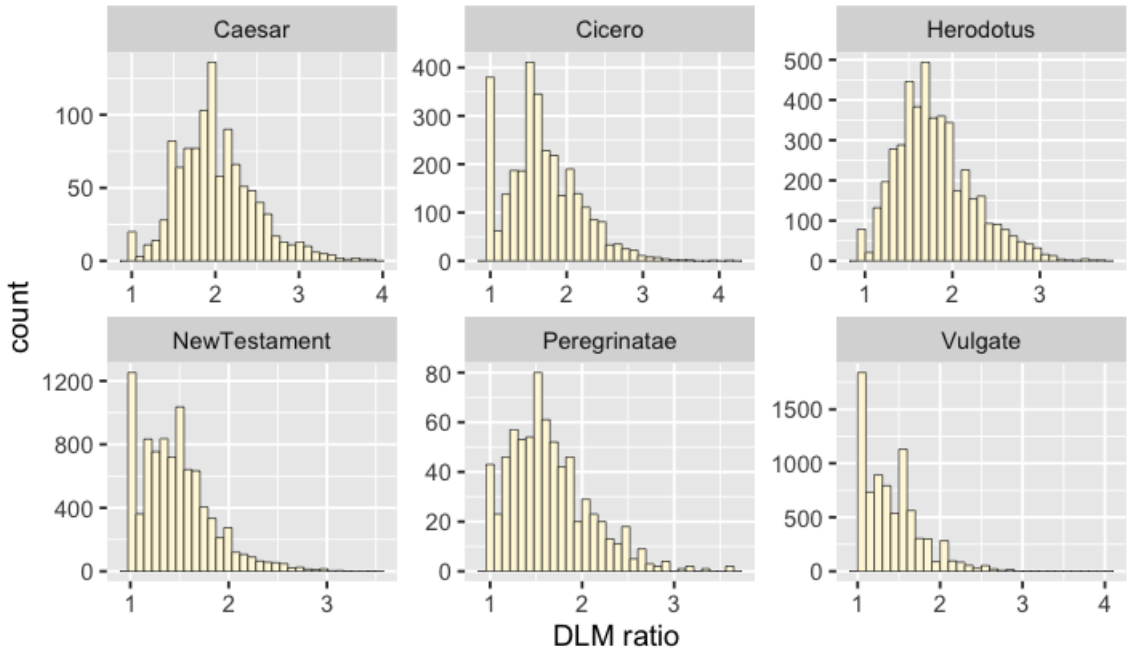


Figure 3.6: Count distribution of DLM ratio values across sentences of all lengths.

tences of length 10 and longer — for the six treebanks in Latin and Ancient Greek. We also report the standard deviation (SD) and standard error (SE) values. Small standard errors confirm, in particular, that the estimation of the DLM ratio is robust. The differences in the DLM ratio between texts in the same language are always larger than twice the highest standard error (0.02) suggesting that their ranking with respect to the degree to which they minimise dependencies is also statistically robust.

The ranking of texts we obtain using the DLM ratio measure is interesting from a diachronic perspective. The texts written in the BC period in Latin (Caesar, Cicero) have higher DLM ratios than the texts written later (Peregrinatio, Vulgate). The same is true for Ancient Greek: Herodotus has a much higher DLM ratio than the New Testament text. This observation is in line with our intuition about diachronic change in Latin and Ancient Greek. These dead languages are known for their notoriously free word order; their descendants — modern Romance and Greek languages — have much more rigid word order. In non-configurational languages, in which words are not necessarily organised in constituents, the freedom of word order is intuitively correlated with longer distances between related words. We should then expect Latin

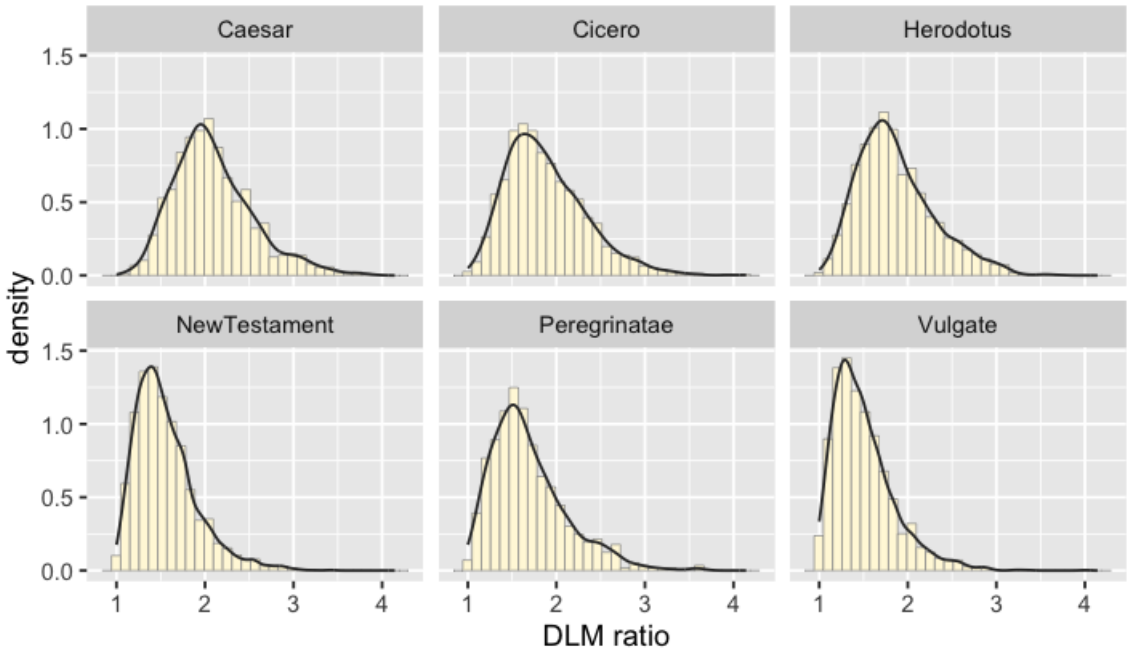


Figure 3.7: Density estimation of the distribution of DLM ratio values across sentences equal to or longer than 10 words.

and Ancient Greek to have progressively shorter dependencies with the loss of word order freedom. Tily (2010) suggests similarly that the minimisation of dependency length in the historical development of English is connected to the loss of case marking and the development of rigid word order. Also, modern Romance languages have very short dependencies as can be seen from the data in Futrell et al. (2015b) and our own results in Section 3.2.

Given the size of our sample — four and two texts in two languages — we cannot claim strong diachronic evidence for DLM. Rather, we take these data as a confirmation that the average DLM ratio is a good treebank-level measure of the degree of DLM as it gives empirical estimations well aligned with our linguistic expectations.

The question raised by the Latin and Ancient Greek data of how DLM is related to word order freedom is not trivial. On the one hand, as we have suggested above, the freedom of word order is intuitively correlated with longer dependency lengths. On the other hand, it is possible in principle that the availability of word order options

Language	Text	# Sent	DLM ratio	SD	SE
Latin	Caesar	752	2.09	0.45	0.016
	Cicero	1256	1.89	0.45	0.013
	Peregrinatio	482	1.70	0.43	0.019
	Vulgate	2271	1.50	0.34	0.007
Ancient Greek	Herodotus	2724	1.86	0.43	0.008
	NewTestament	3568	1.56	0.34	0.006

Table 3.2: Average DLM ratio (for sentences of length 10 and longer), its standard deviation and standard error computed for the PROIEL treebanks.

facilitates the choice of shorter dependencies according to the DLM principle. In addition, we expect that some languages can have longer dependencies than other languages even if all have relatively fixed word order. For instance, a language with the head-final or the head-initial order. The use of the SOV or VSO structure implies that such languages will have longer dependencies between subject and verb, or verb and object, in comparison to SVO languages.

To our knowledge, the lengths of dependencies have not been previously analysed in the typological literature, either as an independent variable or as a correlate of word order freedom. The DLM ratio measure is interesting, therefore, from a typological point of view because it allows us to quantify the degree of DLM in a language and to answer questions such as "Which language minimises dependencies more?" A related typological question is then "Which language has freer word order?" Recently, a number of studies have proposed to quantify word order variability using treebank-based measures (Liu, 2010; Futrell et al., 2015a). In the following, we introduce these measures and apply them to compute word order variability in the PROIEL treebanks.

3.1.3 Measuring word order variability

Very generally, word order variation in a sentence means that a number of different possible grammatical orders of its words can be used to express the same particular syntactic structure. The number and the relative frequency of these orders in production indicate how variable the order in a sentence is. For example, if there is only one grammatical order, then there is no word order freedom at all; the more possible orders there are, the more word order freedom we can attribute to the language. Formally, the extent of word order variation in a sentence can be expressed as the conditional entropy H of the probability distribution of possible orders $o_1, \ldots o_k$ given the words w and the syntactic structure of the sentence t:

$$H(order|w,t) = -\sum_{i=1\ldots k} P(o_i|w,t) \cdot logP(o_i|w,t). \tag{3.3}$$

The probability distribution of orders is taken to be their relative frequency in natural language production. The entropy of this distribution will give us the measure of variability of the order: the higher the number of possible orders k and the more uniform their probabilities, the higher the entropy of the distribution. If only one order is available ($k = 1$), entropy will be minimal and equal to 0.

The overall measure of word order freedom in a language is the entropy of the word order summed over all possible sentences (i.e., their words and syntactic structure):

$$H(order) = \sum_{w,t} p(w,t) \cdot H(order|w,t). \tag{3.4}$$

For a sample of N sentences, the approximation of total entropy is simply the average entropy of all sentences $s_i = (w_i, t_i)$:

$$H(order) \approx \frac{1}{N} \sum_{i=1}^{N} H(order|w_i, t_i). \tag{3.5}$$

A fundamental problem for an accurate estimation of the total entropy is the sparsity of the observed distribution of sentences and their possible orders. In fact, the creative

power of a language is such that it is hard to find the same sentence twice in any corpus or dialogue. We can partly alleviate this problem by conditioning word order variation only on unlexicalised syntactic structures, e.g., unlexicalised dependency trees. Instead of words, we take the parts of speech as the nodes of a dependency tree. However, the combinatorial space of all possible unlexicalised trees still remains unbounded and hard to estimate. Futrell et al. (2015a) investigated an approximation of the total entropy through the factorisation of complete unlexicalised trees into small subtrees consisting of a head and its children. They found that this simplified estimation still has significant sparsity issues and is not robust on samples of 1'000 sentences.

For the reasons outlined above, in this work, we will use a very simple but robust approximation of word order entropy which is the entropy of the direction of dependencies. This measure has been previously employed by Liu (2010) to quantitatively describe the typology of word order in dependency treebanks. Futrell et al. (2015a) have found that the head-direction entropy can be robustly estimated already in samples of as few as 1'000 sentences.

Instead of considering how the complete order of an unlexicalised tree can vary, we consider only how the order between a head and its children varies. For each head h and child c, two word orders — pre-head and post-head — are possible: $o_1 = h\ c$, $o_2 = c\ h$. The entropy of the distribution of these two orders gives us the degree of variability of the direction of dependency $h \rightarrow c$. More precisely, we compute the conditional entropy of dependency direction given the part-of-speech tags of the head h and the child c and the dependency relation r between them: $H(dir|h, c, r)$. The overall entropy is the average arc-direction entropy $H(dir)$ across all dependencies $h \xrightarrow{r} c$ observed in a treebank:

$$H(dir) \approx \frac{1}{N} \sum_{i=1}^{N} H(dir|h_i, c_i, r_i) \tag{3.6}$$

Note that we need to take into account both the functional relation and the part-of-speech tags of the words in a dependency. Functional relations carry crucial information about the syntactic structure of a sentence. For example, a noun and a

verb can appear in both a subject-verb and an object-verb relation. The entropy of these two relations should be computed separately. Part-of-speech tags often carry essential syntactic information in addition to the functional relation. For instance, determiners, numerals and adjectives are all in the modification relation with their head noun. However, they can have different orders with respect to the noun, as is the case in Romance languages, where determiners and numerals precede the noun, but adjectives most often follow the noun. Confounding the various modifier types under one functional relation can lead to overestimation of the entropy.

Arc-direction entropy is sensitive to some but not all aspects of word order variation. It can be used, for instance, to capture the difference between adjective-noun word order properties in Germanic and Romance languages. In English, this word order is fixed, as adjectives appear almost exclusively prenominally; the arc-direction entropy for the adjective-noun dependency will, therefore, be close to 0. In Italian, by contrast, adjectives can both precede and follow nouns; the arc-direction entropy will be greater than 0.

Importantly, arc-direction entropy does not take into account word order variation between sister nodes. For example, the word order variability in the postverbal domain in English does not contribute to the arc-direction entropy value. This aspect of word order variation is already harder to approximate since we need to take into account more parameters of variation with respect to the simple arc-direction entropy (i.e., the head PoS, child_1 PoS, child_2 PoS, relation_1, relation_2). We have conducted several preliminary experiments with sister-order entropy, computed similarly to arc-direction entropy. However, we do not use this measure in this chapter for two reasons. First, the values of sister-order entropy were strongly correlated with the values of the arc-direction entropy.[6] Secondly, as expected, the estimation of the sister-order entropy was more sensitive to the treebank size than the arc-direction entropy. In this work, we focus therefore on the arc-direction entropy measure, which is, at the moment, the most studied and robust approximation of word order freedom. While we believe that developing and analysing measures of word order variation based on treebank statistics is a useful direction for quantitative typological research,

[6]On the fourteen treebanks in the sample used in the parsing evaluation study (Section 3.2), the correlation was equal to 0.87 ($p < 0.001$).

we also suggest an alternative perspective on measuring word order variation in Chapter 5, inspired by the development of an integrated word order production system.

Arc-direction entropy in PROIEL treebanks

The values of arc-direction entropy computed for our six PROIEL treebanks are presented in Table 3.3. These values range between 0.35 (Caesar) and 0.47 (Herodotus). For Ancient Greek, the older Herodotus text has more word order freedom (entropy of 0.47) and longer dependencies (DLM ratio of 1.86) than the more recent New Testament (0.38 and 1.56, respectively). Interestingly, the entropy values do not align perfectly with the DLM ratio and the diachronic scale in Latin: Caesar, for example, has the longest dependencies but lower entropy (0.35) than the more recent Peregrinatio and Vulgate texts, which have higher entropy (0.43) but shorter dependencies (1.5–1.7 against 2.1) . A probable explanation for these results is that arc-direction entropy aggregates many types of word order variation phenomena and is not sensitive enough to track the changes in word order which may affect only some of the constructions. Alternatively, we can speculate that the change in word order is guided by a general tendency to reduce complexity in languages. It is more readily observed in the degree of DLM since dependency lengths are directly related to the processing complexity. On the other hand, word order variation and processing complexity have not been linked directly before. The diachronic development of DLM and word order variation is an intriguing research question but to address it adequately would require a much larger sample of languages.

In addition, our empirical results suggest that the arc-direction entropy and DLM ratio measure capture potentially two different aspects of word order. We confirm this observation in the next section, where we demonstrate that these measures also have a practical application: they serve as useful correlates of parsing performance across typologically diverse languages.

Language	Text	Entropy	DLM ratio
Latin	Caesar	0.35	2.09
	Cicero	0.43	1.89
	Peregrinatio	0.43	1.70
	Vulgate	0.43	1.50
Ancient Greek	Herodotus	0.47	1.86
	New Testament	0.38	1.56

Table 3.3: The arc-direction entropy values (Entropy) computed for the Latin and Ancient Greek treebanks in our sample. DLM ratio values computed previously are given for comparison.

3.2 Evaluating the effect of word order properties on parsing performance

The measures we have presented and assessed qualitatively in the previous section constitute a potentially interesting way to quantify word order properties at the treebank level. These can be used, for example, to describe and compare the properties of languages in a typological study. In this section, we will show that measures of word order properties at the treebank level can also be used to inform multilingual NLP technology. Specifically, we will apply the DLM ratio and entropy measures to evaluate the effect of corresponding word order properties on parsing performance.

Parsing is the task of producing a correct syntactic analysis of a sentence given its surface representation — a string of words. Modern statistical parsing systems are supervised machine learning algorithms which are given a treebank with sentences and their manual syntactic annotation as training data. After training on these so-called gold trees, they perform very well on unseen sentences; for instance, some recent systems achieved correct labelled dependencies above 92% for English (Chen and Manning, 2014; Andor et al., 2016). With the development of treebanks for many other languages, there is a growing interest in building multilingual parsing systems

that can be trained and can obtain good results in any language without changes to the parsing architecture. The results for many languages — in particular, for languages with relatively free word order and rich morphological systems — are typically inferior to the results on English (Seddah et al., 2011, 2013; Andor et al., 2016; Zeman et al., 2017). In fact, it is a common assertion that high word order variability and longer dependencies negatively affect parsing performance. English has a relatively fixed order and short dependencies. Its word order is unambiguously defined for many syntactic relations such as modifier-noun or verb-object, which simplifies syntactic analysis in some respects. Evaluations of the effect of dependency length on parsing performance in English have been conducted by Rimell et al. (2009); Nivre et al. (2010). McDonald and Nivre (2011) used 13 languages from the CONLL-X shared task (Buchholz and Marsi, 2006) and analysed the effect of sentence and dependency lengths on the parsing performance averaged across these languages. Unfortunately, a cross-linguistic large-scale analysis of the effect of word order properties is not straightforward. Multiple confounding factors affecting parsing performance hamper such an analysis. Apart from few basic factors such as the size of the training treebank, these confounding factors are very hard to control for. For example, different treebanks will have different average sentence lengths, different lexicon sizes and different percentages of words in the test set that were never observed in the training set. Probably even more crucial for cross-linguistic parsing evaluation is the difference between annotation schemes. Indeed, this is one of the main reasons behind the development of Universal Dependencies treebanks (Agić et al., 2015; Nivre et al., 2016, 2017).

In this section, we evaluate the effect of word order properties on parsing performance, measured as the DLM ratio and arc-direction entropy. First, we will confirm that longer dependencies are harder to parse on the treebanks of Latin and Ancient Greek. These treebanks constitute a special evaluation set-up since they contain several texts that come from the same language but from different time periods and, presumably, differ minimally in their properties such as the lexicon and morphology compared to the word order. Secondly, we propose a new framework for artificially creating treebanks that are minimal pairs with respect to the word order properties of interest, similar to the set-up of PROIEL treebanks. This framework allows us to

perform an analysis of word order in parsing across many languages with different morphological and syntactic properties.

3.2.1 Background: Dependency parsing and evaluation

Since we work with dependency treebanks, we perform the evaluation of dependency parsing systems (as opposed to constituency parsing systems). In this section, we briefly describe the statistical approach to dependency parsing, including the architecture of MaltParser, which we use in our experiments. We also examine previous dependency parsing work that analyses the effect of word order properties and dependency length on parsing performance.

Dependency parsing: basic notions

Dependency parsing consists of producing a dependency tree representation t given an input sentence s. Statistical dependency parsing is based on learning an underlying function $f : s \rightarrow t$ given a gold treebank, i.e., a set of pairs $\{(s_i, t_i)\}$. Statistical dependency parsing received growing attention in the NLP field, starting with the work of Nivre et al. (2006) and McDonald et al. (2006). At the moment, dependency parsing is the most prominent parsing subfield thanks to the expansion of parsing technology to languages other than English and the development of multilingual dependency treebanks.

The two main architectures used for dependency parsing are the so-called transition-based and graph-based architectures. In this work, we conduct our experiments using a popular system known as MaltParser, which is a canonical implementation of the transition-based architecture Nivre et al. (2006). Dependency parsing is being constantly improved. While we present the results only for MaltParser, the evaluation framework we propose can be used to evaluate and analyse the performance of all types of dependency parsers.

Evaluation of a parser trained on one portion of a treebank is conducted on a distinct, test portion of a treebank. We will use the notation $T_{train} \rightarrow T_{test}$ to refer

to an evaluation scenario with training data T_{train} and test data T_{test}. The measures standardly used for dependency parsing evaluation are unlabelled and labelled accuracy scores (UAS and LAS). UAS computes how many dependencies $w_i \rightarrow w_j$ output by a parser are found in the gold dependency annotation. LAS computes how many labelled dependencies $w_i \rightarrow^l w_j$, including the functional label l of the dependency, are produced correctly.

MaltParser

MaltParser is an implementation of a stack-based transition parsing architecture. It belongs to a more general bottom-up (or shift-reduce) type of parser that constructs a parsing tree starting from its leaf nodes and tries to gradually combine them into pieces of syntactic structure. Transition-based parsers process the leaf nodes — the words in a sentence — from left to right in an incremental fashion.

More formally, instead of learning to construct a syntactic tree directly, the parser is trained to construct its derivation incrementally, constrained by the input string of words and a set of possible actions at each step. A derivation consists of transitions between states. Possible transitions can include, for example, attaching a newly observed word to a previously observed word, creating, therefore, a dependency between them. Parsing architecture defines which transitions a parser can use and how it encodes a partially analysed sentence in its intermediate configurations. Shift-reduce parsers use a buffer to encode the unobserved part of a sentence and a stack to encode the part which was already observed and partially analysed. The most basic set of transitions includes left-arc and right-arc reduce actions (create a dependency between the word on top of the stack and the first word in the buffer) and a shift action (move the first word of the buffer to the stack).[7] This set of transitions is sufficient to produce derivations for all projective dependency trees.

In addition to the set of possible transitions, the parsing algorithm needs to specify how one of these transitions is chosen at each step in the derivation. For data-driven parsers, the best transition is identified using a statistical classifier. The classifier is

[7]For a more accurate, formal description of transitions, see, for example, Kübler et al. (2009).

trained on the gold derivations extracted from the training treebank and learns which transition to use in which configuration. Importantly, the configuration is encoded in terms of features such as "the part of speech of the word on top of the stack" or "the number of dependents previously attached to the word on top of the stack". The performance of MaltParser is affected significantly by the choice of the features; in this work, we use MaltOptimiser (Ballesteros and Nivre, 2012) to automatically identify the best feature set based on held-out validation data. MaltParser also implements several variations in the parsing algorithm which employ different sets of transitions, including those required for deriving non-projective dependency trees. Based on validation data, MaltOptimiser chooses the most appropriate of these algorithms for the treebank.

Effect of word order properties on parsing performance

The work of Rimell et al. (2009) and several follow-up experiments with dependency parsers (Nivre et al., 2010; Bender et al., 2011; Merlo, 2015) analyse parsing performance in syntactically complex constructions involving long dependencies in English. These constructions include subject and object relative clauses, wh-questions and other constructions characterised by a long dependency between a verb and one of its arguments. Because of the recursive structure of the language, the length of such dependencies is, in principle, unbounded. Nivre et al. (2010) found that the two representatives of dependency parsers — transition-based MaltParser and graph-based MSTParser — perform much worse on these hard constructions than on average in a treebank. Note that the constructions investigated in these studies are very infrequent: the most frequent subject relative clause construction appears in only 6 to 10% of sentences; other constructions appear in only 0 to 3% of sentences (Rimell et al., 2009). Consequently, the test set containing these constructions is small — 560 sentences — and was extracted semi-automatically using the treebank annotation. Since this construction-focused parsing evaluation methodology is language-specific and can require manual extraction and evaluation, it is not surprising that it has not been extended to other languages.

A more large-scale evaluation of parsing performance with respect to the lengths

of dependencies is presented in McDonald and Nivre (2011). It is also conducted only in English, but the lengths of dependencies are analysed for all sentences and construction types. The results are presented for the two main-stream dependency parsers: MaltParser and MSTParser. First, the study establishes that both parsers have declining performance with growing sentence length. According to the authors, this is primarily due to the increase in the presence of complex syntactic constructions such as the ones studied in the work discussed above. The accuracy of parsing is also shown to decrease for longer dependencies. In addition, McDonald and Nivre (2011) look at other structural factors that can affect parsing performance, such as the tree depth (the distance to the root, in their notation) and the branching factor (i.e., the number of siblings). An increase in either of these parameters affects parsing performance negatively. The results of the evaluation experiments in McDonald and Nivre (2011) are used to inform a better parsing model combining the advantages of both MaltParser and MSTParser. Note that, on the basis of the reported results, we cannot establish an explicit relation between any of the parameters studied and the parsing performance, in particular, because all the parameters are correlated. In this work, we aim for a parsing evaluation analysis which allows us to control for sentence length and syntactic structure and to evaluate the effect of lengths of dependencies independently of these correlated factors.

To our knowledge, no previous studies have addressed in a similar systematic manner other word order properties such as the degree of word order variation in a language. The most relevant observations come from the Shared Tasks on morphologically-rich languages (Seddah et al., 2011, 2013). The tasks are set up to evaluate both constituent and dependency parsers and provide detailed treebank information related to lexical and morphological complexity such as the size of the lexicon, the average number of word types per token and similar metrics. A comparison across nine languages in a dependency parsing scenario with the same training set size (5'000 sentences) shows, for example, that Korean and Hebrew ($\sim$ 83% LAS) treebanks are harder to parse than French or Polish ($\sim$ 89% LAS) treebanks. However, even if the training size is the same, it is not clear whether Korean and Hebrew are harder because of their morphological richness and associated word order freedom or because of other factors such as the average sentence length in the treebank, the size of the dependency

label set and so on. The organisers of the shared task perform some very general correlation analyses, e.g., comparing the ratio of the treebank size and the size of the label set against the parsing performance, and admit that they "cannot tell why certain treebanks appear more challenging to parse than others, and it is still unclear whether the difficulty is inherent on the language, in the currently available [parsing] models, or because of the annotation scheme and treebank consistency" (Seddah et al., 2013, p.175).

3.2.2 Parsing evaluation on Latin and Ancient Greek treebanks

Given the previous evidence of the effect of long dependencies for parsing, we can put forward the following hypothesis on DL at the sentence level:

> *A dependency tree with a small overall dependency length should be easier to parse than a tree with a large overall dependency length.*

To test this hypothesis at the treebank level, we can use the DLM ratio to measure the extent to which the dependency trees minimise the overall dependency length. If there is a systematic relation between the length of dependencies and the parsing performance, treebanks with a lower DLM ratio should be easier to parse than treebanks with a higher DLM ratio. Given the PROIEL treebank collection, which contains several texts in the same language annotated with the same annotation scheme, we have an opportunity to test this hypothesis on texts that constitute the controlled minimal pairs for such analysis.

To evaluate the effect of word order properties and to verify that the texts in the same language but from different epochs share the same lexicon and can be considered minimal pairs, we test several training and testing configurations. Specifically, we use two different set-ups: training and testing within the same text and across texts of different periods.

First, we evaluate the parsing performance across time periods. As training data, we use texts from one period with similar word order properties (e.g., two texts in Latin from the BC period) and test the parsing performance on texts from a different

Language	Training treebank	Test treebank	Train Size	UAS
Latin	BC	AD	67k	67.27
	AD	BC	106k	57.72
Ancient Greek	Herodotus	New Testament	75k	76.05
	New Testament	Herodotus	120k	61.27

Table 3.4: Parsing accuracy for period-based training and test configurations for Latin and Ancient Greek.

period (e.g., two texts in Latin from the AD period). For both Latin and Ancient Greek, we perform therefore two evaluations: BC $\rightarrow$ AD and AD $\rightarrow$ BC. We evaluate the parsing performance using the unlabelled accuracy score (UAS). The results of these four evaluation configurations are presented in Table 3.4. We can observe that the results are relatively high, allowing us to conclude that the training and test texts are indeed written in the same language and share a lexicon and other properties that are crucial for the adequate performance of the parser. Despite the uncontrolled training size of the data for the BC $\rightarrow$ AD and AD $\rightarrow$ BC scenarios, we can note that parsing of BC texts — which have longer dependencies and generally higher word order variation according to our DLM ratio and entropy measures (Table 3.3) — is harder than parsing of AD texts (57.7 UAS vs 67.3 UAS for Latin and 61.3 UAS vs 76.1 UAS for Ancient Greek). In fact, since the training data in the AD $\rightarrow$ BC scenarios is larger than in the BC $\rightarrow$ AD scenarios, we can safely attribute the difference in performance to the difference in word order properties.

To further confirm our result in a more controlled setting, we perform a set of evaluations using training and test data from the same text. For the "within-text" evaluation, we apply a standard random split, with 90% of the corpus assigned to training and 10% assigned to testing, for each text separately. We eliminated potentially confounding effects due to different training sizes by including only around 18'000 words for each text in Latin (the size of the Peregrinatio corpus), and around 75'000 in Ancient Greek. The results of these experiments are given in Table 3.5.

Language	Treebank	Training size	UAS	Entropy	DLM ratio
Latin	Caesar	18k	66.46	0.35	2.09
	Cicero	18k	63.11	0.43	1.89
	Peregrinatio	18k	74.35	0.43	1.70
	Vulgate	18k	83.92	0.43	1.50
	all texts	155k	78.30		
Ancient Greek	Herodotus	75k	69.76	0.47	1.86
	New Testament	75k	88.01	0.38	1.56
	all texts	195k	79.94		

Table 3.5: Parsing accuracy for random-split training (90%) and test (10%) configurations for each language and for each text independently. The entropy and DLM ratio values are duplicated from Table 3.3.

In general, we can observe that the older Latin and Ancient Greek texts have lower UAS scores than their more recent counterparts which have more fixed word order with shorter dependencies. Interestingly, Caesar has slightly higher parsing accuracy (66.5 UAS) than Cicero (63.1) which can be due to its lower arc-direction entropy (0.35 versus 0.43) which counteracts the higher DLM ratio (2.1 versus 1.9).

As a side result, we also report a strong baseline for each language, calculated by training and testing on all texts combined and split randomly with a 90%/10% split. The cumulative parsing accuracy on both Latin and Ancient Greek is relatively high as seen from the 'all texts' random split configuration. These performance values are especially high compared to the previous results reported for other Latin and Ancient Greek treebanks, e.g., in LDT and AGDT with 61.9% and 70.5% of UAS, respectively (Lee et al., 2011). This increase in accuracy is likely due to the fact that our texts are prose and not poetry.

To summarise, in both across-text and single-text experiments, we see that the accuracy for older texts written in Latin in the BC period is much lower than the accuracy for late Latin texts written in the AD period. This pattern correlates with the

previously observed smaller degree of dependency length minimisation and word order variation of BC texts compared to AD texts. Similarly, for Greek, Herodotus is much more difficult to parse than the New Testament text, which corresponds to their differences in the rate of DLM as well as the entropy of word order. Our results provide the first confirmation obtained in a controlled evaluation setting for the general assertion that freer order languages are harder to parse. As we stressed in this section, the collection of Latin and Ancient Greek treebanks presents an especially advantageous set-up for such evaluation. In general, we cannot directly correlate the performance of a parser on two different languages and treebanks with their DLM and entropy values since there are many other confounding properties of the treebanks. Also, the PROIEL data do not allow us to separate the effects of the two word order properties which are partially correlated. To address these challenges, we propose a new framework to artificially construct minimal pair treebanks for analysis of word order properties.

3.2.3 Creating artificial treebanks for minimal pair evaluation

General methodology

The new evaluation methodology we propose consists in modifying an existing treebank T to create an artificial treebank T', so that T' is its minimal pair with respect to some property of interest, and analysing the parsing performance by comparing the results on the two treebanks. In this section, which focuses on analyses of word order properties in parsing performance, we create several kinds of artificial treebanks in the same manner: each sentence s' in T' is a permutation of the words of the original sentence s in T. We permute words in various ways according to the word order property whose effect on parsing we want to analyse. Crucially, we change only the order of the words in a sentence; the dependency tree structure t of a permuted sentence s' in T' always remains the same as in the original sentence s in T. Formally, if $T = \{(s_i, t_i)\}$ then $T' = \{(s'_i, t_i)\}$, where $s' = permutation(s)$.

Given a permutation, for each treebank in our sample of languages we conduct two parsing evaluations: $T_{Train} \rightarrow T_{Test}$ and $T'_{Train} \rightarrow T'_{Test}$, where the training-test data

split for T and T' is always the same, that is, $permutation(s) \in T'_{Train} \iff s \in T_{Train}$ and $permutation(s) \in T'_{Test} \iff s \in T_{Test}$. As before, the parsing performance is measured as unlabelled and labelled attachment scores (UAS and LAS), the proportion of correctly attached arcs in the unlabelled or labelled tree, respectively.

Given the training–testing set-up, the differences in unlabelled attachment scores $\Delta UAS = UAS(T_{Test}) - UAS(T'_{Test})$ can be directly attributed to the differences in word order properties o between T and T', setting aside other treebank properties h. More formally, we can assume that $UAS(T) = f(o^T, h^T)$ and $UAS(T') = f(o^{T'}, h^T)$. Except for word order properties o^T and $o^{T'}$, the two equations share all other treebank properties h^T — such as the size of the treebank, its average dependency length, the size of PoS tagset — and f is a function that applies to all languages, here embodied by a given parser.

This methodology can also be used to analyse parsing performance at the sentence level. Consider a pair of sentences s and its permuted variant s'. The two sentences share all lexical items and underlying dependencies between them. Consequently, if the parsing accuracy on the two sentences is not the same, the explanation must be sought in their different word orders. In standard treebank evaluation settings, exact sentence-level comparisons are not possible, as two sentences very rarely constitute a truly minimal pair with respect to any specific syntactic property. Our approach opens up the possibility of a deeper understanding of parsing behaviour at the sentence level and even of individual dependencies based on large sets of minimal pairs.

Permutations to test DLM and arc-direction entropy

We create two types of permuted treebanks to optimise for the two word order parameters — dependency length and word order variability — which we showed can be measured robustly as the DLM ratio and arc-direction entropy. We perform two types of word order permutations to the treebanks in our sample: a permutation that minimises the lengths of the dependencies in a dependency tree and a permutation that minimises the variability of word order. Below we describe the permutation

procedures in more detail. For each original treebank and its permuted versions, we compute the DLM ratio and arc-direction entropy values, as described in the section 3.1. We then compare how the parsing performances on the original and the permuted trees vary in relation to the differences in the DLM ratio and entropy across the pairs of treebanks.

Creating trees with minimal DL Given a sentence s and its dependency tree t in a natural language, we employ the algorithm proposed by Gildea and Temperley (2010) to create a new artificial sentence s' with a permuted order of words. The algorithm reorders the words in a sentence s to yield a projective dependency tree with the minimal overall dependency length $DL(s')$.[8] To do so, it recursively orders the subtrees consisting of a head and its immediate children, starting from the root node. For each subtree, first, all children nodes are sorted according to the number of nodes in the subtree headed by the child node (including all its descendants), i.e., according to the length of the phrase. The algorithm then places the children ordered in this way $c_1, c_2, \ldots$ on the left and on the right of the head in alternation, starting from the shortest child phrase. The children on the same side of the head are also ordered based on their sizes, with the shortest phrases closer to the head, producing, e.g., a resulting order such as $\ldots c_5 c_3 c_1 h c_2 c_4 \ldots$. Children of the same size are ordered between each other as found in the original sentence.

Note that this algorithm is deterministic and that the dependency length of each sentence is optimised independently. By definition, the DLM ratio for sentences permuted in such a way is equal to 1. As we will see, this type of permutation generally leads to very high arc-direction entropy in the treebank, since the order of children with respect to their head is not constrained by the grammar in any way.

We exclude from our analysis sentences with any non-final punctuation marks and sentences with multiple roots. In natural language treebanks, punctuation marks such as commas or parentheses are typically attached to the head of the clause or

[8] In principle, an order with minimal DL can be non-projective. However, such cases are rare in natural language trees, which have limited topology. In particular, natural language trees have small average branching factors, while a non-projective order with minimal DL occurs only if at least one node of out-degree 3 is present in the tree (Chung, 1984).

the sentence they appear in. Often this creates a long dependency when punctuation marks serve to separate two phrases or indicate their edges (in a parenthetical use). These long dependencies do not determine the degree of DLM in a language from a linguistic perspective. However, the frequency of punctuation marks can affect the value of the DLM ratio and bias our results. We decided, therefore, to exclude sentences with punctuation from our analysis (excluding the period which was simply removed for all sentences).[9]

Creating trees with minimal entropy To obtain treebanks with a minimal arc-direction entropy equal to 0, we can fix the order of each type of dependency, defined by a tuple (rel, h, c). There exist therefore many possible permutations resulting in zero arc-direction entropy. We choose to assign the same direction (either Left or Right) to all the dependencies. This results in two permutations yielding fully right-branching (RB) and fully left-branching (LB) treebanks. We order the children on the same side of a head in the same way as in the OptDL permutation: the shortest children are closest to the head. For the RB permutation, children of the same size are kept in the order of the original sentence; for the LB permutation, this order is reversed, so that the RB and LB orders are symmetrical. These two permutations are particularly interesting, as they give us the two extremes in the space of possible tree-branching structures. Moreover, since the LB/RB word orders for each sentence are completely symmetrical, the two treebanks constitute a minimal pair with respect to the tree-branching parameter.

Importantly, there exist both predominantly right-branching (e.g. English) and left-branching natural languages (Japanese, Persian) and the comparison of LB- with RB-permuted treebanks will show how much of the difference in parsing typologically different natural languages can be attributed to their different branching directions. Of course, the parsing sensitivity to the parameter depends on the parsing architecture.

A transition-based parser such as MaltParser relies on left-to-right processing of words and the fully right-branching or fully left-branching orders can yield potentially different results.

[9]Note that Latin and Ancient Greek texts do not have punctuation marks. As a result, this problem has not arisen in our previous experiments.

3.2.4 Experiments with MaltParser on 14 treebanks

Parsing set-up

For all our experiments we use MaltParser, introduced in section 3.2.1. For optimal outcomes, the transition-based MaltParser must be provided with a list of features tailored for each treebank and each language. We use the MaltOptimizer package Ballesteros and Nivre (2012) to find the best features based on the training set. We conduct four training-test evaluations per gold treebank: using its original natural language sentences and OptDL-permuted, LB-permuted and RB-permuted variants. The training-test splits are identical for the four evaluation scenarios.

Dependency treebanks

We use a sample of 14 dependency treebanks for 12 languages. The treebanks for Bulgarian, English, Finnish, French, German, Italian and Spanish come from the Universal Dependencies project and use the same annotation scheme (Nivre et al., 2016). We use the treebank for Dutch from the CONLL 2006 shared task (Buchholz and Marsi, 2006). The Polish treebank is described in Woliński et al. (2011) and the Persian treebank in Rasooli et al. (2013). For comparison, we add two Latin and two Ancient Greek dependency annotated texts from the PROIEL collection Haug and Jøhndal (2008). We include the Cicero and Vulgate texts in Latin and the Herodotus and New Testament texts in Ancient Greek. The quantitative properties of these treebanks are presented in Table 3.6 (second and third column). This set includes treebanks that had at least 3'000 sentences in their training set after elimination of sentences not fit for permutation (with punctuation marks or multiple roots). This excluded from our analysis some otherwise typologically interesting languages such as Basque and Arabic. Where available, we used the training-test split of a treebank provided by its distributors; in other cases, we split the treebank randomly with a 9-to-1 training-test set proportion.

Language	Abbr.	Size	Av. sentence length	DLM ratio	Entropy
Italian	it	57k	12.1	1.30	0.18
Spanish	es	63k	15.1	1.32	0.16
French	fr	72k	14.5	1.32	0.12
Polish	pl	29k	6.8	1.33	0.36
Bulgarian	bg	30k	8.5	1.36	0.20
English	en	62k	9.5	1.40	0.10
Finnish	fi	46k	5.7	1.42	0.35
Vulgate (La)	la.V	63k	8.8	1.50	0.43
NewTestament (AG)	el.NT	69k	10.5	1.56	0.38
Dutch	nl	38k	8.4	1.62	0.29
German	de	65k	11.5	1.65	0.22
Herodotus (AG)	el.H	59k	14.4	1.87	0.47
Cicero (La)	la.C	35k	11.6	1.88	0.44
Persian	fa	35k	9.4	1.99	0.16

Table 3.6: Training size (in number of words), average sentence length, DLM ratio and arc-direction entropy (Entropy) measures for the treebanks in our sample.

Word order properties of original and permuted treebanks

Table 3.6 also presents the values of the DLM ratio and entropy measures calculated on the training set of the original non-permuted treebanks. From these data, we confirm that the DLM ratio and entropy measures capture different word order properties as they are not correlated (Spearman correlation $r = 0.38$, $p > 0.1$). For example, we can find languages with a low DLM ratio and high entropy (Finnish) and with a high DLM ratio and low entropy (Persian). Furthermore, these two measures do not necessarily reflect genetic similarity between languages of the same family; for example, two languages from different language families — Polish (Indo-European family) and Finnish (Finno-Ugric family) — are situated close to each other in the space of two word order parameters.

Table 3.7 shows how the DLM ratio and entropy values change when we apply the two permutations to the treebanks. Compared to the values of the original treebanks, the DLM ratio and entropy values of the artificial treebanks are much more narrowly distributed: 1.35 ± 0.05 (mean $\pm$ standard deviation) compared to 1.54 ± 0.24 for the DLM ratio and 0.65 ± 0.02 compared to 0.28 ± 0.13 for entropy. Importantly, the treebanks in the LB-/RB-permuted set have, on average, both lower entropy and a lower DLM ratio than the original treebanks. The treebanks in the OptDL set have a lower DLM ratio but higher entropy than the original treebanks. We expect these differences in DLM and word order variability measures to affect parsing performance evaluated using UAS and LAS for each of the four sets of treebanks. More precisely, as outlined in Section 3.2.3, we assume that the difference in UAS values (e.g., between the original and LB-permuted treebanks) — $\Delta UAS = UAS(LB) - UAS(Original)$ — will depend on the difference in the DLM ratio — $\Delta DLMratio = DLMratio(LB) - DLMratio(Original)$ — and the difference in entropy — $\Delta Entropy = Entropy(LB) - Entropy(Original)$ (computed from Table 3.7). Smaller values of delta measures indicate shorter dependencies and less variable word order and should lead to better parsing performance in terms of larger ΔUAS values. Below, we present the parsing performance measures obtained for our 14 treebanks and test these assertions empirically.

Language	DLM ratio			Entropy		
	Original	OptDL	LB/RB	Original	OptDL	LB/RB
Italian	1.30	1.00	1.34	0.18	0.65	0.00
Spanish	1.32	1.00	1.36	0.16	0.66	0.00
French	1.32	1.00	1.38	0.12	0.66	0.00
Polish	1.33	1.00	1.35	0.36	0.62	0.00
Bulgarian	1.36	1.00	1.36	0.20	0.65	0.00
English	1.40	1.00	1.38	0.10	0.65	0.00
Finnish	1.42	1.00	1.42	0.35	0.64	0.00
Vulgate (La)	1.50	1.00	1.34	0.43	0.65	0.00
NewTestament (AG)	1.56	1.00	1.33	0.38	0.66	0.00
Dutch	1.62	1.00	1.22	0.29	0.58	0.00
German	1.65	1.00	1.44	0.22	0.67	0.00
Herodotus (AG)	1.87	1.00	1.38	0.47	0.67	0.00
Cicero (La)	1.88	1.00	1.31	0.44	0.65	0.00
Persian	1.99	1.00	1.32	0.16	0.66	0.00
$\text{Mean}_{(\pm\text{st. deviation})}$	$1.54_{\pm 0.24}$		$1.35_{\pm 0.05}$	$0.28_{\pm 0.13}$	$0.65_{\pm 0.02}$	

Table 3.7: The DLM ratio and arc-direction entropy (Entropy) measures for the original and permuted treebanks in our sample. The two 'LB/RB' columns present the measures for LB-/RB-permuted treebanks optimised for zero entropy; the two 'OptDL' columns present the measures for treebanks optimised for the minimal DLM ratio.

Results

Table 3.8 presents the parsing performance values of MaltParser for all treebanks and permutation scenarios.

Overall, all three sets of permuted data are easier to parse than the original data. We observe an increase of +1% and +6% UAS for OptDL and LB/RB data, respectively. The better results on the LB-/RB-permuted data must be due to the property of the treebank highlighted previously: the LB/RB data have both lower entropy and a lower DLM ratio than the original data.

The performance of the parser on our artificial treebanks confirms that the lengths of the dependencies and the word order variability are two factors that negatively affect parsing accuracy. Two illustrative examples are the texts in Latin, a language with highly variable word order, and German, a language known for its long dependencies (as confirmed by its high DML ratio of 1.65). For the Cicero text, for example, we can conclude that its variable word order is indeed the primary reason for the very low parsing performance (67% UAS). These numbers improve significantly when the treebanks are rearranged in a fixed LB/RB word order (88% UAS). This permutation reduces the DLM ratio by 0.57 and reduces entropy by 0.44, yielding a very considerable increase in UAS of 21%. The other permutation, which minimises dependency lengths, reduces the DLM ratio by 0.88 but increases entropy by 0.21. This increase in entropy dampens the beneficial effect of DL reduction, and performance increases 12%, less than in the fixed-order permutation. For German, our analysis gives the same overall results. The DLM ratio in the RB/LB scenario decreases slightly (from 1.65 to 1.44) and its entropy also decreases (-0.22). The performance of the parser on RB-/LB-permuted data is better than on the original data (89% versus 86% UAS). Moreover, when the DLM ratio is reduced (-0.65, in the OptDL permutation), but entropy is increased (from 0.22 to 0.67), we find a reduction in performance (from 86% to 84% for UAS). These data suggest that the word order variability of German, minimised in the RB/LB case, has a potentially higher impact on parsing difficulty than its long dependencies.

A more detailed picture emerges when we compare pairwise the original treebanks

Language	Original		OptDL		LB		RB	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
it	93.9	90.6	90.7	84.6	94.2	90.0	95.0	90.8
es	85.7	80.7	80.1	71.7	85.2	76.0	87.8	80.3
fr	84.1	80.0	81.3	73.7	90.0	82.4	90.7	85.2
pl	92.1	88.0	94.3	88.4	93.7	89.2	93.6	88.9
bg	92.6	88.6	91.6	84.5	91.9	85.1	92.7	87.1
en	89.6	87.7	84.7	78.7	89.3	82.9	88.8	83.4
fi	82.9	79.7	85.2	80.6	90.1	84.7	90.7	86.6
la.V	86.0	80.5	87.6	80.9	92.5	85.9	92.5	86.4
el.NT	84.8	79.0	88.1	80.7	92.5	85.0	90.5	73.0
nl	88.4	84.3	92.8	87.0	95.0	89.9	94.7	89.9
de	85.5	79.7	83.7	75.0	88.8	77.9	89.3	81.2
el.H	71.7	65.0	83.0	73.7	88.8	79.3	87.6	66.7
la.C	67.4	58.8	78.6	67.4	87.6	75.8	87.2	76.1
fa	82.7	73.7	83.5	73.3	89.8	79.7	89.7	79.9
Average	84.8	79.7	86.1	78.6	90.7	83.1	90.8	82.5

Table 3.8: Parsing performance results measured as unlabelled and labelled accuracy scores (UAS and LAS, %) for four types of treebanks in 14 languages: original treebanks, their versions permuted for minimal dependency length (OptDL) and their versions permuted for minimal arc-direction entropy (LB/RB).

to the permuted treebanks for each of the languages. For this analysis, we use only the UAS measure, since attachment decisions are more directly dependent on word order than labelling decisions, which are mediated by correct attachments. Hence, we limit our analysis to three parameters: the DLM ratio, entropy and UAS.

Figures 3.8 (OptDL) and 3.9 (LB) plot the differences in UAS of MaltParser between pairs of the permuted and the original treebanks for each language to the differences in DLM ratio and entropy between these treebanks. Our dependent variable $\Delta UAS = UAS(T') - UAS(T)$ is computed from Table 3.8. The x-axis and y-axis values $\Delta DLM = DLMRatio(T) - DLMRatio(T')$ and $\Delta Entropy = Entropy(T) - Entropy(T')$ compute the differences of the measures between the original treebank and the permuted treebank based on the numbers in Table 3.7. Note that we have chosen to calculate these differences reversing the two factors, compared to the ΔUAS value, for better readability of the figures: an increase in the entropy or dependency length values corresponds to a decrease in the difficulty of parsing and, therefore, to the increase of the dependent variable ΔUAS.

For the OptDL data (Figure 3.8), the overall picture is coherent with the previously observed patterns: the more the DLs are minimised and the less entropy is added to the artificial treebank, the larger the gain in parsing performance (violet-to-blue circles in the lower left corner and yellow-to-red circles in the upper right corner). Again, we observe an interaction between the DLM ratio and entropy parameters: for the languages with a relatively low DLM ratio and low entropy originally, such as English or Spanish, the performance on the permuted data decreases. This is because, while the DLM ratio decreases, entropy increases. For this group of languages, the particular trade-off between these two properties leads to lower parsing accuracy.

The LB-permuted data show similar trends (Figure 3.9). An interesting regularity is shown by four languages (Latin Vulgate, Ancient Greek New Testament, Dutch and Persian) on the off-diagonal. Although they have different relative entropy and DLM ratio values, which span from near minimal to maximal values, the improvement in parsing performance on these languages is very similar (as indicated by the same colour). This again strongly points to the fact that both the DLM ratio and entropy contribute to the observed parsing performance values.

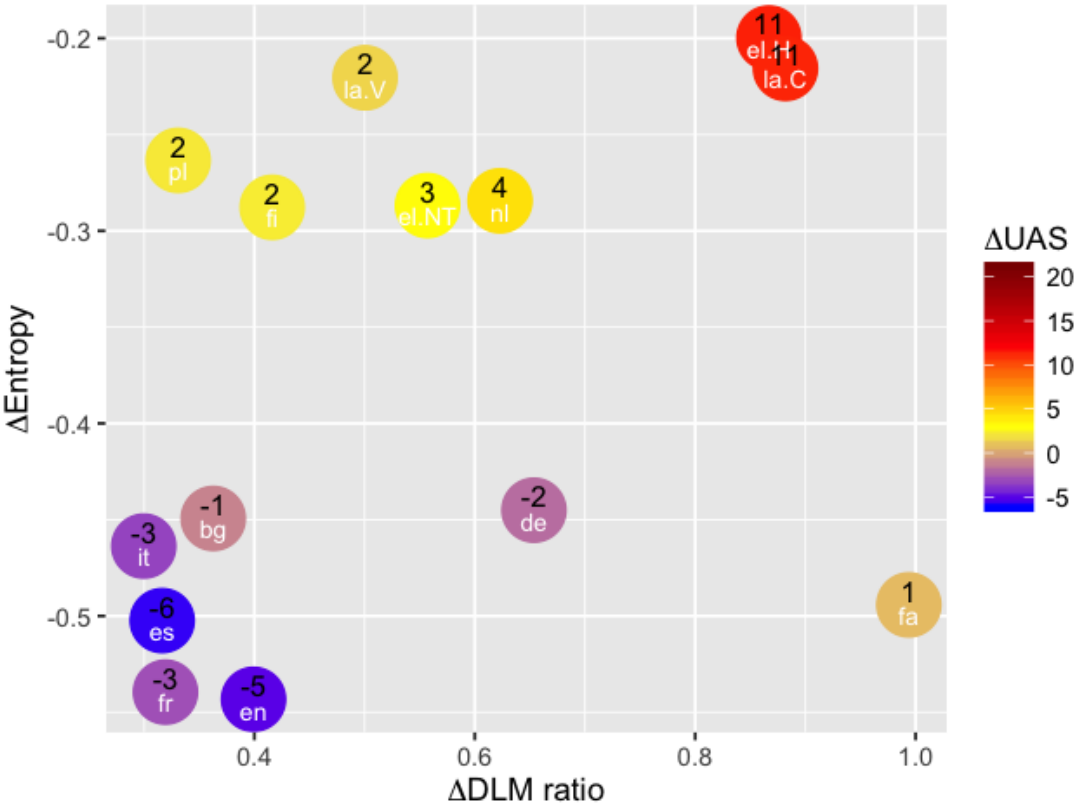


Figure 3.8: Differences in UAS of MaltParser between OptDL-permuted and original pairs of treebanks for the corpora in our sample.

We can further confirm the effect of dependency length by comparing the parsing accuracy across sentences.[10] Consider the Dutch treebank and its RB-permuted pair. For each sentence and its permuted counterpart, we can compute the difference in their dependency lengths ($\Delta DLM = DLM - DLM_{RB}$) and compare it to the difference in parsing performance ($\Delta UAS = UAS_{RB} - UAS$). We expect to observe that ΔUAS increases when ΔDLM increases. Indeed, the parsing results for Dutch show a positive correlation between these two values ($r = 0.40, p < 0.001$). Note that this sentence-level monolingual analysis is different from the similar analysis of the effect of dependency length on parsing performance in English (McDonald and Nivre, 2011) in an important way. In particular, the DLM ratio measure is independent of

[10]Note that the entropy measure is computed on a whole treebank and cannot be meaningfully compared across sentences.

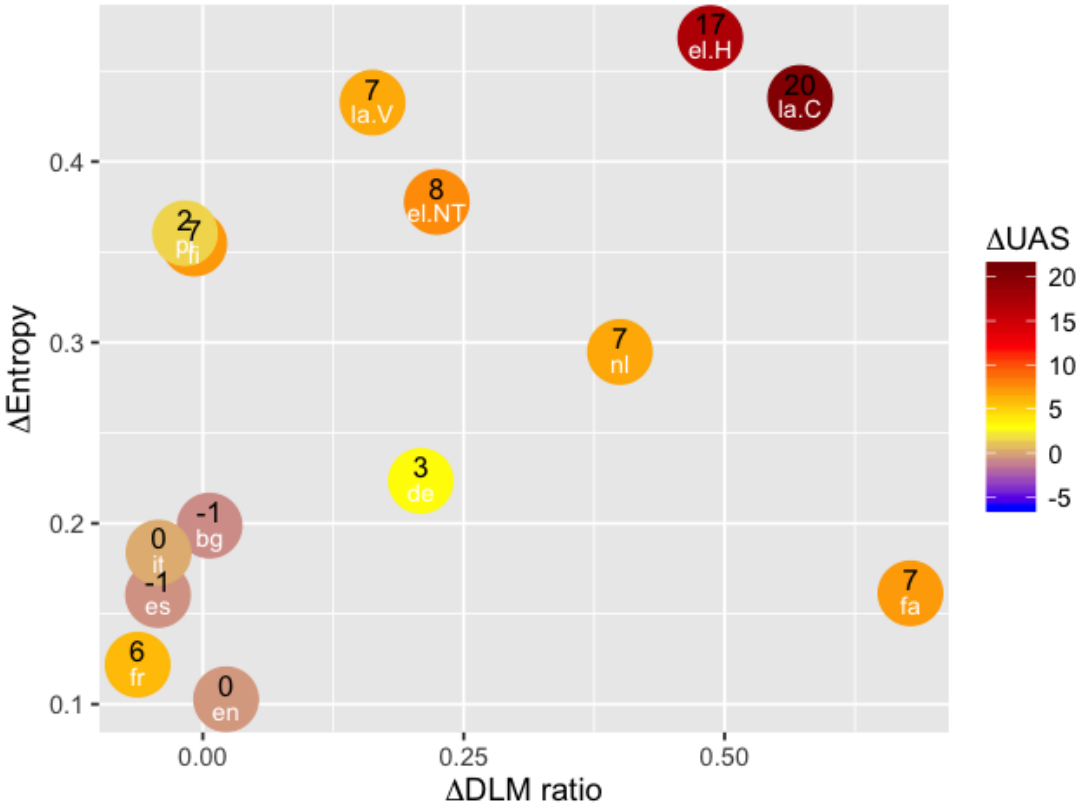


Figure 3.9: Differences in UAS of MaltParser between LB-permuted and original pairs of treebanks for the corpora in our sample.

the sentence length, and we can, crucially, separate the effect of longer dependencies from the effect of longer sentences.

All these analyses confirm and quantify that dependency length and word order variability affect parsing performance.

Sentence-level analysis of parsing performance

Looking at Table 3.8, we observe that MaltParser shows the same average accuracy for RB- and LB-permuted data. However, some languages show significantly different results between their LB- and RB-permuted data, especially in their labelled accuracy scores. The New Testament corpus, for example, is much easier to parse when it is

rearranged in left-branching order (91% RB vs 93% LB UAS, 73% RB vs 85% LB LAS). Our artificial data allows us to investigate this difference in the scores by looking at parsing accuracy at the sentence level.

The differences in MaltParser accuracies on RB- and LB-permuted data are striking because these data have the same head-direction entropy and dependency length properties. The only word order difference is in the branching parameter, resulting in two completely symmetrical word orders for each sentence of the original treebank. To understand the behaviour of MaltParser, and of transition-based parsers in general, we looked at the out-degree, or branching factor, of the syntactic trees. The intuition is that when many children appear on one side of a head, the parser behaviour on head-final and head-initial orders can diverge due to sequences of different operations, such as *shift* versus *attach*, that must be chosen in the two cases.[11]

The data for the New Testament treebank indicates that the branching factor plays a role in the differences between LB and RB parsing scenarios. For each pair of sentences with LB/RB orders, we computed the parsing accuracies (UAS and LAS) and the branching factor as the average out-degree of the dependency tree. We then tested whether the better performance on the LB data is correlated with the branching factor across the sentences ($UAS_{LB} - UAS_{RB} \sim BF$). The Pearson correlation for UAS values was 0.08 ($p = 0.02$), but for LAS values the correlation was 0.30 and highly significant ($p < 0.001$). On sentences with larger branching factors, the labelled accuracy scores on the LB data were higher than on the RB data.

We combine our result for the branching factor with an observation based on the confusion matrix of the labels, to provide a more accurate explanation of the comparatively low LAS in the RB-permuted treebank of the New Testament corpus. We found that when a verb or a noun has several one-word children, such as 'aux' (auxiliaries), 'atr' (attributes), 'obl' (obliques), or 'adv' (adverbs), these elements frequently receive the wrong label if they appear after the head (RB data), but are labelled correctly if they appear before the head (LB data). It appears that the leftward placement of children is advantageous for the transition-based MaltParser: at the moment of the

[11] The MaltParser configurations for LB and RB data had the same parsing algorithm (Covington projective).

first attachment decision for the child closest to the head, it has access to a larger left context. When children appear after the head, the first child is attached before any other children are seen by the parser and the labelling decision is less informed, leading to more labelling errors.

It should be noted that it is not always possible to identify a single source of difficulty in the error analysis. Contrary to the New Testament, Spanish is easier to parse when it is rearranged into the right-branching order (88% RB vs 85% LB UAS, 80% RB vs 76% LB LAS). However, the types of difficult dependencies emerging from the different branching of the LB/RB data were not similar or symmetric to those of New Testament. In the case of Spanish, we did not observe a distinct dimension of errors that would explain the 4% difference in UAS scores.[12]

3.2.5 Perspectives for parsing evaluation using artificial treebank data

Our results highlight both the contributions and the challenges of the proposed evaluation framework. On the one hand, the results show that we can identify and manipulate word order properties of treebanks to analyse the impact of these properties on parsing performance and suggest avenues to improve it. In this respect, our framework is similar to standard analyses of parsing performance based on separate manipulations of individual word-level features (such as omitting morphological annotation or changing coarse PoS tags to fine PoS tags). Similarly to these evaluation procedures, our approach can lead to improved parsing models or a better choice of parsing model by discovering their strengths and weaknesses.

In addition to MaltParser, we also evaluated MSTParser using the same framework. The results of this evaluation were reported in the TACL paper (Gulordava and Merlo, 2016). MaltParser and MSTParser are not directly comparable due to differences in the training set-up (MaltParser features are optimised for each language and

[12] Overall, the variance in the LB/RB performances on Spanish is relatively high, and the mean difference (computed across UAS scores for sentences) is not statistically significant (t-test: $p > 0.5$) – a result we would expect if errors cannot be imputed to clear structural factors.

permutation). Nevertheless, MSTParser performs slightly better on average than MaltParser on permuted datasets. Also, MSTParser does not show any difference in performance on LB-/RB-permuted treebanks. In general, analysing several parsers in the same evaluation framework advances our understanding of parsing architectures. Subsequently, when two parsing systems are known to have different strengths and weaknesses, they can be successfully combined in an ensemble model for more robust performance (Surdeanu and Manning, 2010; McDonald and Nivre, 2011).

A contribution of the parsing performance analyses in a multilingual setting is the identification of difficult properties of treebanks. For the Cicero and Herodotus texts, for example, our method reveals that their word order properties are causes for the low parsing performances compared to the other languages. This result confirms our linguistic intuition, but it could not be formally concluded without factoring out confounds such as the size of the training set or the dissimilarity between the training and test sets, which could also be reasons for low parsing performance. Together, the knowledge of word order properties of a language and the knowledge of parsing performance related to these properties give us an a priori estimation of which parsing system could be better suited for a particular language.

On the other hand, our method also raises some complexities. Compared to commonly used parsing performance analyses related to word-level features, the main challenges to a systematic analysis of word order lie in its multifactorial nature and in the large choice of quantifiable properties correlated with parsing performance. The multifactorial nature of word order means that it is very hard to manipulate one word order property in isolation from other word order properties. The two properties we have looked at — the DLM ratio and arc-direction entropy — cannot be manipulated independently since minimising one property leads to the increase of the other. Another challenge is due to the fact that, as we have seen in Section 3.1, formal quantitative approaches to studying word order variation cross-linguistically are just beginning to appear and not all word order features have been robustly quantified.

Our method, which consists in creating artificial treebanks, can prove useful beyond parsing evaluation. For instance, our data could enrich the training data for tasks such

as de-lexicalised parser transfer (McDonald et al., 2011). Word order properties play an important role in computing similarity between languages and finding the source language leading to the best parser performance in the target language (Naseem et al., 2012; Rosa and Zabokrtsky, 2015). A possibly large artificially permuted treebank with word order properties similar to the target language could then be a better training match than a small treebank of an existing target natural language. Shortly after the publication of our work (Gulordava and Merlo, 2016), a paper which explores exactly this idea by constructing very many artificially permuted treebanks was published (Wang and Eisner, 2016).

3.3 Conclusions

This chapter demonstrates how we can analyse DLM effects and word order variation at the language-level by quantifying these word order properties in dependency-annotated treebanks. We argued that the DLM ratio — the average ratio between the dependency length of a sentence and its minimal possible dependency length — is a robust measure for comparison of the degree of DLM across languages and treebanks. Measuring word order variation as a unique parameter is a harder task because of the sparsity of the observed combinations of the order of words. The arc-direction entropy captures robustly one aspect of word order variation: the variability in the position of a child with respect to its head. Our empirical analysis of PROIEL and other treebanks suggests that DLM ratio and arc-direction entropy provide information about orthogonal or complementary properties of word order. The measures of DLM and word order variation at the language level can, therefore, give important new information for typological comparisons of the languages of the world.

In addition to linguistic studies, the measures of word order at the treebank level can be useful for natural language processing applications. We successfully applied these measures to analyse, quantitatively and on a large scale, the effect of word order properties on parsing performance. We proposed two new scenarios for a controlled evaluation: the comparison between texts in the same language (Latin and Ancient Greek) using PROIEL treebanks and the comparison between minimal pair

treebanks constructed artificially by permuting the order of words in sentences. Using these evaluation scenarios, we confirmed that treebanks and sentences with longer dependencies are harder to parse while controlling for many confounding factors (including, for example, the sentence length). We also confirmed experimentally, for the first time to our knowledge, a common intuition that treebanks with higher word order variation are harder to parse.

The experiments we have presented in this chapter highlight clearly that there is large space for future work on exploring and quantifying word order properties at the language level. We investigated only a restricted set of properties which, of course, do not describe word order exhaustively. New interesting data could come from quantifying other word order properties and studying the properties we have touched upon in much more detail and across many more languages. A very exciting direction for future research lies in understanding and quantifying the relation between DLM and word order freedom.

Parsing performance is also affected by syntactic and word order properties other than those we have studied in this chapter including, for instance, the branching factor or the position of the root node. Luckily, our artificial treebank evaluation framework allows us to test many potential properties in the same systematic way without running into the problem of data sparsity. In fact, while we conducted our experiments on only three types of permuted treebanks, we can easily construct many more permutation types. Each permuted treebank with specific word order parameter values provides us with a data point in the evaluation space. Thus, we can obtain, in principle, many data points for a statistically reliable estimation of the effects of word order properties, even if we investigate several properties at the same time.

Chapter 4

DLM effects in adjective-noun order variation

This chapter investigates the dependency length minimisation (DLM) effects in the variation of prenominal and postnominal placement of adjectives in Romance languages. We pursue a traditional corpus-based multifactorial analysis to test the effects of lengths of dependencies on the variation. Our aim is to verify whether the general DLM principle is at work in the adjective-noun variation — a construction which has not received a lot of attention in relation to DLM. By probing the variation in this way, we find some properties in the distribution of adjectives which have not been pointed out before. In this new perspective, a number of syntactic phenomena previously treated separately, such as the postnominal preference of heavy adjectives and the effect on the adjective placement of the dependents in a complex noun phrase, can be explained in terms of one principle.

4.1 Background

4.1.1 Adjective variation in Romance and heavy adjectives

Adjectives modifying nouns in Romance languages have two possible positions: preceding the noun or following it, as illustrated in the examples (4.1–4.4) for a noun phrase *a difficult situation* in Italian, French, Spanish and Portuguese:

(4.1) Italian:
  a. una difficile situazione
  b. una situazione difficile

(4.2) French:
  a. une difficile situation
  b. une situation difficile

(4.3) Spanish:
  a. una difícil situación
  b. una situación difícil

(4.4) Portuguese:
  a. uma difícil situação
  b. uma situação difícil

The *default position* of adjectives in all Romance languages is uncontroversially assumed to be the **postnominal** position (variant **b** in the examples above). In fact, it is the most frequent order overall in the corpus-based data and, when asked to choose between the two minimal pair word orders as in (4.1–4.4), without the sentence context, native speakers have strong preferences for the postnominal order. In such simple noun phrases consisting only of an adjective and a noun, speakers could also judge the prenominal order to be ungrammatical. The acceptability judgments and production preferences are influenced noticeably by the context in which the noun phrase appears in the sentence. For a slightly modified noun phrase *a difficult*

economical situation (instead of just *a difficult situation*), the prenominal word order (i.e., *une difficile situation economique* in French) is already more acceptable than the order **a** in the examples (4.1–4.4).[1]

Despite the many potential factors affecting the placement of adjectives in a sentential context, the investigation of this construction in theoretical syntax has mostly focused on simple noun phrases of the type (4.1–4.4) and the semantic differences between the prenominal and postnominal adjectives which present a puzzle on their own. First, some classes of adjectives such as adjectives of color and nationality can appear only postnominally (in all Romance languages):

(4.5) a. una camicia rossa
b. *una rossa camicia
'a red shirt'

(4.6) a. un ragazzo americano
b. *un americano ragazzo
'an American boy'

Secondly, there exist few adjectives in each language which exhibit robust alternations in their meaning when used prenominally and postnominally:[2]

(4.7) a. un pauvre homme
'a pitiful man'
b. un homme pauvre
'a broke man'

(4.8) a. une grande actrice
'a great actress'
b. une actrice grande
'a tall actress'

[1] Based on the informal queries of native speakers, p.c. The corresponding frequency patterns in production can be verified through a corpus or Google search queries.
[2] Thuilier (2012) lists, for example, nine such adjectives in French referring to them as homophones.

Finally, adjectives can have the same lexical meaning but different semantic interpretation when used prenominally and postnominally. This alternation has received special attention in the literature (Bouchard, 1998; Alexiadou, 2001; Truswell, 2005; Cinque, 2010).[3] One of the differences in the interpretation corresponds to the restrictive versus non-restrictive scope of the adjective over the noun.

(4.9) Spanish (example from Alexiadou (2001)):

a. el oloroso lirio (non-restrictive)

b. el lirio oloroso (restrictive)

'the fragrant lily'

(4.10) Italian:

a. le strette strade (non-restrictive)

b. le strade strette (restrictive)

'the narrow streets'

A postnominal adjective (4.9b, 4.10b) specifies a quality of the noun which distinguishes its referent from a set of nouns of this type (*el lirio oloroso* is the one lily which is fragrant; *le strade strette* is a subset of all streets). A prenominal adjective, by contrast, denotes a presupposed, non-restrictive quality of the noun. Similarly, the postnominal position is often associated with adding the contrast or establishing the difference ('the narrow streets and not the wide ones') while the prenominal position is more neutral and provides an attributive characterization to the noun ('some streets that happen to be narrow'). In a number of extreme theoretical accounts, the position of an adjective is isomorphic to its semantic interpretation (Waugh, 1977; Bouchard, 1998; Cinque, 2010). It implies that *all* adjectives appearing prenominally and postnominally always receive different interpretation in the two positions.

We adopt here an alternative view on the relation between the syntax and the semantics of adjective position. In our opinion, the account of variation in adjective

[3] In fact, this topic has been of much interest for French linguists starting already from 18th century (Roubaud (1786)). This early work is reviewed in Waugh (1977) and Forsgren (1978). See also Truswell (2005) and Blöhdorn (2008) for a brief summary.

placement based only on the differences in restrictive and non-restrictive interpretation cannot be taken as a complete explanation for preferences between prenominal and postnominal orders. It is recognised that the slight differences in interpretation are not robust: they are not easily identified and acknowledged by all native speakers (Abeillé and Godard, 1999; Thuilier, 2012). This is particularly true for subjective attributive adjectives such as *difficult, interesting, charming, horrible* etc. For instance, there is no difference in the interpretation of the two orders in the example (4.11) in French.

(4.11) French (example from Abeillé and Godard (1999)):

a. un jeune homme *charmant*

b. un *charmant* jeune homme

'a charming young man'

In this work, we will assume that, for this type of adjectives that can appear both prenominally and postnominaly, the two word order variants have the same meaning. More specifically, we suggest that, for the noun phrases which make part of naturally occurring sentences (the kind of noun phrases which we will be analysing) the semantic factors such as the ones proposed by, e.g., Bouchard (1998) and Cinque (2010), do not play the first role in defining the adjective placement preferences. In addition to lexical constraints (such as the obligatory postnominal position for adjectives of color), other syntactic, discourse and phonological properties can play an important role in the adjective-noun variation. They have been analysed in the previous literature but to a lesser extent. The work presented in this chapter extends this previous work and provides new empirical data on adjective variation in sentential context through a cross-linguistic corpus-based study of five Romance languages.

Specifically, this chapter is dedicated to a systematic analysis of the effect of syntactic properties in the noun phrase related to dependency length minimisation. We survey below the literature directly relevant to this aspect of adjective variation. We first describe the phenomenon of heavy adjective shift, potentially related to the DLM principle, and the work that has addressed this variation. We then discuss the

previous corpus-based statistical work on the variation in adjective placement, in particular, the extensive analysis of French data by Thuilier (2012).

Heavy adjective shift

In English, while the normal position of an adjective is prenominal, an adjective phrase with a complement should appear obligatorily after the noun (4.12). The same postnominal requirement for adjectives with a complement holds also for Romance languages (4.13).

(4.12) a. a man [proud of his achievements]
b. *a [proud of his achievements] man

(4.13) Italian translation of (4.12):
a. un uomo [orgoglioso delle sue riuscite]
b. *un [orgoglioso delle sue riuscite] uomo

A similar restriction was observed by Greenberg (1963) for adverbial modifiers of the adjectives in a typological sample of languages: the order Adj Adv N is not grammatical (contrasting with other possible orders Adv Adj N, N Adv Adj and N Adj Adv) as illustrated for English in the example (4.14).

(4.14) a. *a running smoothly meeting
b. a smoothly running meeting

Williams (1982) generalises these data under the Head-Final Filter principle — a constraint preventing post-head material in prenominal modifiers. In other words, he postulates an adjacency requirement for the head of the parent phrase (the noun) and the head of the modifier phrase (the adjective). When a complement or an adverb

of the adjective intervenes between the adjective and the noun, such order is not acceptable.[4]

Languages can employ different ways to avoid this dispreferred order, for example, by extraposing the dependent of the adjective (4.15) or by placing the whole adjective phrase postnominally, which is the default order in Romance languages (4.16).

(4.15) a *difficult* book [for anyone to read]

(4.16) un libro [*difficile* da leggere per chiunque]

However, as Abeillé and Godard (2000) note, there are many counter-examples to Williams' generalisation. For example, some adverbs in English can appear in Adj Adv N order as in *a fair enough proposal*. Also, in some languages such as Russian, the [Adj PP] N order is at least marginally acceptable (4.17).

(4.17) [гордый до слез] Иван[5]

[proud up to (his) tears] Ivan

Abeillé and Godard (2000) propose for French an alternative principle based on the heaviness of adjective phrases. In addition to the data presented above, this principle is devised to account for the other data on the variation in adjective position such as the tendency of many adjective phrases (non-bare adjectives) to appear postnominally in French. Abeillé and Godard (2000) suggest that this tendency is conditioned on the syntactic feature *[lite]/[non-lite]* and that the prenominal and postnominal adjectives must be, respectively, *lite* and *non-lite*. For bare adjectives, the value of the feature is defined in the lexicon. For example, the postnominal-only adjectives of color are specified as *non-lite* and therefore occur postnominally. The feature is also defined for other lexical elements such as adverbs. For an adjective phrase, the weight value is a

[4] The Head-final Filter was shown to be connected to a more recent theoretical attempt at describing typological patterns of variation as a Final-over-Final constraint in the grammar (Sheehan, 2017). This principle, in turn, makes predictions similar to the Hawkins' domain minimisation processing preferences (Sheehan, 2012).

[5] Interestingly, the prenominal adjective order seems to be more acceptable when the noun phrase is a sentence final subject, e.g. *Их встретил [гордый до слез] Иван* (gloss: 'them met proud up to tears Ivan').

result of the syntactically-informed combination of the values of the feature of the phrase elements. If a *lite* adjective is modified by an adverb which is *lite* (e.g., *très* 'very') then the adjective phrase will be also *lite*. If, instead, the adverb is non-lite (e.g., *politiquement* 'politically') then the adjective phrase will also be *non-lite* and will appear postnominally. Adjectives with (post-head) complements are always non-lite and therefore cannot appear prenominally, in line with the Head-Final Filter.

This proposal for "heavy" (*non-lite*) adjectives to appear postnominally is reminiscent of the similar treatment of noun phrases and their rightward extraposition often referred to as heavy-NP shift (Section 2.2). It was observed, for example, that a noun with a relative clause can be extraposed more easily than a simple noun phrase (e.g. with only prenominal modifiers) (Wasow, 2002).

This connection between the "heavy adjective shift" and heavy-NP shift appears also in empirical studies of adjective distribution in French (Forsgren, 1978; Thuilier, 2012). In fact, Forsgren (1978) proposes explicitly that the short-before-long principle applies in the case of the noun-adjective pair and results in the longer adjective phrases preferring postnominal placement (observed in his corpus study). As we will show in the next section, while in the case of postverbal complements, the short-before-long principle is directly equivalent to dependency length minimisation, it is not clear whether this is true for the case of adjective-noun pair.

We can obtain some preliminary evidence in favor of the adjective shift generalisation from a frequency analysis of a small number of languages whose treebanks are readily available (Figure 4.1). The data are based on a sample of languages from the UD treebank collection and show the percent of postnominal placement for two categories — simple (bare) adjectives (green bars) and heavy adjectives (adjective phrases, red bars). Based on this coarse distinction, we observe that there are more postnominal heavy adjectives compared to postnominal simple adjectives. However, this evidence is very preliminary and does not necessarily suggest a more general DLM principle at work (as the work on head-final languages suggested for the alternation of verbal dependents cross-linguistically). We cannot say therefore whether the heavy adjective postposition preference can be connected to a more general case of dependency length minimisation, as it is the case of heavy-NP shift in English. We leave a rigorous cross-

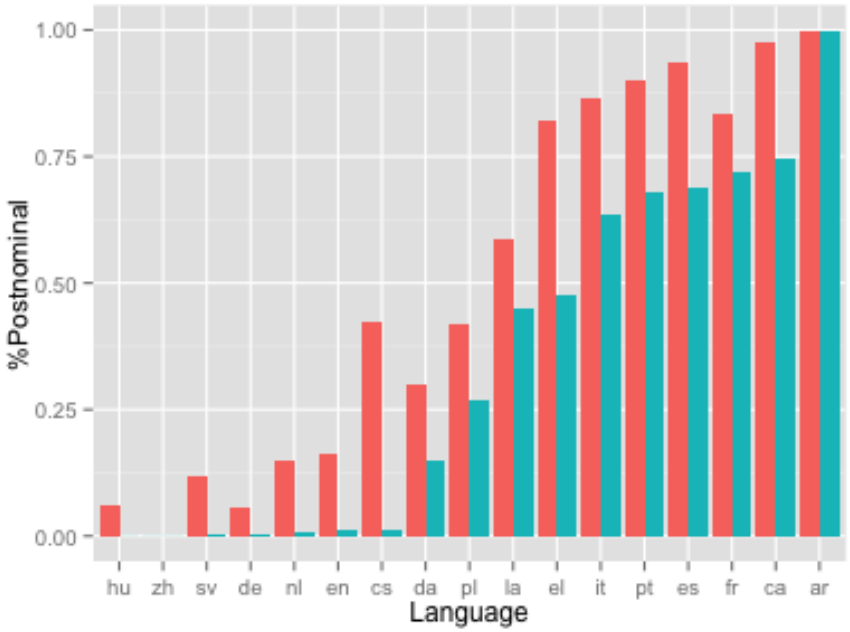


Figure 4.1: The illustration of difference between percetage of postnominal simple (green bars) and heavy (red bars) adjectives across several languages.

linguistic analysis of the adjective-noun variation beyond the Romance languages for future work.

In this chapter, we focus, instead, on the distribution of adjectives in complex noun phrases in Romance and show that these data can already provide new insight on whether the DLM principle is at work in adjective-noun variation and heavy adjective shift. We propose a theoretical formalisation of DLM for the case of the noun phrase and conduct the empirical statistical tests of the DLM predictions on the corpora of five Romance languages.

Corpus-based analysis of adjective variation in French

A rather small number of quantitative corpus-based studies on the order of adjectives were conducted for English on the relative order between several prenominal adjectives (Wulff, 2003) and Romance languages on the relative order between adjective and noun (Waugh, 1977; Forsgren, 1978; Centeno-Pulido, 2010; Fox and Thuilier, 2012;

Thuilier et al., 2012). We are not aware of any cross-linguistic corpus-based studies of adjective variation for more than one Romance language. We describe here in detail the work of Thuilier (2012) which is the most complete statistical treatment of adjective position in one of the Romance languages – French — and is very related to our work from the methodological point of view.

The work described in Thuilier et al. (2012) and Fox and Thuilier (2012) on adjective placement in French provides an extensive analysis of a large number of factors affecting the patterns of variation. This work follows a corpus-based empirical approach to variation based on the same methodology that the work of Gries (2003) and Bresnan et al. (2007) described previously. Similarly, the goal of this work is to assess the relative effect of the factors previously proposed in the literature using naturally occurring data and a multifactorial statistical model.

The factors analysed include lexical factors such as whether the adjective and the noun form a collocation, semantic factors such as the class of the adjective (an adjective of nationality or color, or a relational adjective), phonological factors (length of the adjective in syllables) and a number of syntactic factors. The syntactic variables, which are the most relevant ones for this overview, cover the structure of the adjective phrase as well as the noun phrase. To test the data on heavy adjective shift presented in the work of Abeillé and Godard (2000), Thuilier (2012) includes the presence of an adverb or a coordination of adjectives in the adjective phrase as variables in the model to verify that they favor the postnominal placement.[6] In addition, the structural composition of the noun phrase such as presence of other noun dependents have been suggested in the previous literature to affect the adjective placement (Forsgren, 1978). In particular, the reference grammar of French (Grevisse and Goosse, 2007) suggests that in cases where there are several modifiers of a noun they should be placed, if possible, on the opposite sides of the noun, to make a more "balanced" noun phrase. For example, in the presence of a relative clause or a prepositional phrase (which always occur after the noun), the adjective should be placed before noun. Finally, the definiteness of the noun phrase (indicated by an article) and the

[6]While the factors such as the presence of adverb or coordination in the adjective phrase are analysed, the presence of adjectival complements is not included in the model. This is because adjectives with complements have to follow obligatorily the noun.

syntactic function of the noun phrase (subject, object or an attribute) were proposed to affect the adjective placement. All these factors were included in the statistical analysis to verify their impact in a large corpus and in one mutlifactorial model.

Several logistic regression models with the independent variables listed above were trained on the corpus data to predict the observed variable: the prenominal or the postnominal placement of an adjective. Simple logistic models were used to analyse the distribution of all adjectives in the corpus and a generalised mixed effect model was used to analyse only the distribution of alternating adjectives. The latter model uses adjective lemmas as random effects, an approach that we pursue and describe in more details in this chapter.

The output of the statistical analysis of adjective placement in French with respect to the syntactic factors is the following. The presence of an adverb or a coordination are significant factors which favor the postnominal placement, confirming thus the heavy adjective analysis proposed by Abeillé and Godard (2000). The presence of a prepositional phrase or a postnominal adjective favor prenominal placement of the target adjective. However, the modification of a noun by a relative clause complement does not carry a significant effect, contrary to the modification by a PP. If the NP is introduced by a definite article (also a possessive or a demonstrative) the prenominal adjective placement was found to be preferred. The syntactic function on the NP, on the other hand, was not found to have an affect on the adjective placement.

The work of Thuilier establishes several results which motivate the experiments we present in this chapter. Most importantly, the analyses of adjective placement were conducted on a syntactically-annotated corpus with mostly automatic encoding of token and sentence features. It was not possible to identify automatically from the corpus the pragmatic and discourse factors (for example, the new or given status of the noun phrase, intonation and so on). Consequently, they were not included in the model. Despite this, the model with all other features (lexical, syntactic, semantic) achieved very high performance in predicting the position of the adjectives (92.6% accuracy for all adjectives and 87% for alternating adjectives). We can conclude from this result that adjective order alternation is mostly constrained by the lexico-semantic properties of the adjective and the syntactic properties of the noun phrase. These

factors can be successfully studied in a corpus without taking into account pragmatic and discourse factors which do not play a main role in the variation. Another important result is the fact that, as discussed above, a number of syntactic factors were shown to play an important role in the adjective placement in addition to the preferences defined in the lexicon.

There remains, however, a number of open questions and issues not addressed in the work of Fox and Thuilier (2012); Thuilier et al. (2012); Thuilier (2012). While the results confirm the appropriateness of the multifactorial statistical approach to explaining the adjective alternation, the interaction of various factors is not taken into account. A preference which ties two or more factors together cannot be expressed in terms of a simple linear combination of variables. From a practical, modelling point of view, it can be infeasible to test for interactions between all pairs of variables. In fact, this consideration leads to a more general limitation of the previous work. A statistical analysis can test and evaluate the effect of factors in a variation but it does not explain *why* these factors are relevant. Thuilier (2012) does not attempt to explain why certain syntactic factors turn out to be significant in her analysis. An important question is whether there is a more general underlying principle that can explain the effect of these syntactic factors together. A theoretical analysis would also be necessary to put forward the potential interactions between the variables which could be then tested using statistical tools and corpus data.

4.1.2 Statistical models for word order variation analysis

Linear regression analysis

A simple linear regression model is a statistical model for two variables, X and Y. The underlying assumption is that there exist a linear relation between X — the predictor variable, also called independent variable — and Y, the target or response variable: $Y = \beta_0 + \beta_1 X$. We can observe and reconstruct this relationship only through a sample of data points $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ which are obtained empirically and are not necessarily drawn from the analytical equation describing Y given X. More precisely, the simple linear regression model is specified by the Equation (4.1), where

X and Y are random variables. A data set is assumed to be sampled from the model given the random noise (or random error) ϵ "corrupting" the measurement of Y.

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{4.1}$$

Given a data set of observations, the goal is to estimate the parameters β_i of the model (also called regression coefficients). We will refer to β_0 as the *intercept* and β_1 as the *slope* of the model. The estimated parameters can then be used to predict the values of Y given the observed values of X and to analyse and test the effect of X on Y.

The maximum likelihood estimation of parameters — finding β_i which maximise the total probability of the observed data $P(Y \mid X, \boldsymbol{\beta})$ — is straightforward for simple linear regression models if the random noise ϵ is assumed to be normally distributed. More precisely, if $\epsilon \sim N(0, \sigma^2)$ then the probability of observing a value y given a value x is:

$$p(Y = y \mid X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-(\beta_0+\beta_1 x))^2}{2\sigma^2}} \tag{4.2}$$

The maximum likelihood estimation amounts to finding the values of β_0 and β_1 which maximise the total probability of the observed data set:

$$\beta_0^*, \beta_1^* = arg\max \prod_{i=1}^{n} p(Y = y_i \mid X = x_i) \tag{4.3}$$

In the case of normally distributed random error, maximum likelihood estimation is equivalent to least squares estimation. Intuitively, it corresponds to finding the regression line which minimises the sum of squared distances from the the dataset points (x_i, y_i) to the predicted points (x_i, y_i*) on the line.

The simple regression model can be generalised to more than one predictor variables $X_1, \ldots X_k$. The *multiple regression model* is specified similarly as:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon \tag{4.4}$$

And can be written in a short form using vector representation:

$$Y = \boldsymbol{\beta X} + \epsilon \tag{4.5}$$

Linear regression is used often in psycholinguistics to study, for example, the relation between the reading time of a sentence (response variable) and its syntactic properties (predictor variables). The chief purpose of such analyses is to evaluate and test the hypotheses associated with the variables X_i (e.g., an object relative clause requires longer processing) on the dependent variable Y (reading time is a proxy measurement of processing effort). Statistically, this amounts to testing whether the coefficient β_i for the variable X_i is different from 0. If this coefficient is equal to 0 we would conclude that there is no (linear) relation between X_i and Y. If β_i is different from 0 it captures whether an increase in the value of X_i results in an increase or a decrease of the value of Y (graphically seen as a positive or negative slope of the line) and by how much. The coefficients β_i are estimated based on the observed data sample and a statistical test is necessary to determine whether β_i is different from 0 in the actual population from which this sample is extracted. More precisely, we want to test the null hypothesis that β_i is equal to zero and possibly reject it based on a computed statistic at some significance level. In case of a multiple linear regression with normally distributed error the t-test statistic is used to check individual parameters β_i.

Logistic models

Linear regression is an appropriate model for continuous response variables such as reading time. In corpus-based syntactic studies, instead, the response variables are normally of categorical type. When modelling word order variation the response variable is most often *binary*: e.g., split or adjacent order in the verb-particle construction (as is modeled in Gries's work described before) or prenominal versus postnominal order of an adjective (as used in the work of Thuilier). To describe the relation between the predictor variables and the binary response variable we use a type of generalised linear model, known as *logit* model.

A binary response variable Y can take two values: 0 and 1. The logistic regression can be seen intuitively as finding the parameters $\boldsymbol{\beta}$ so that:

$$Y = \begin{cases} 1, & \text{if } \boldsymbol{\beta X} + \epsilon > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.6}$$

Under the assumption that random error is distributed by the standard logistic distribution, we can then write the probability of $Y = 1$ as follows:

$$p(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{e^{\boldsymbol{\beta x}}}{1 + e^{\boldsymbol{\beta x}}} \tag{4.7}$$

Alternatively, the same relation can be expressed using the inverse of the logistic function — the logit transform:

$$\boldsymbol{\beta x} = \ln \frac{\mu}{1 - \mu}, \tag{4.8}$$

where $\mu = p(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$.

The estimation of parameters in logit models cannot be obtained from analytical computation similarly to the maximum likelihood estimates for the linear regression models. Instead, numeric estimation algorithms are typically employed and implemented in statistical software such as R language. The significance of individual coefficients β_i is assessed by computing the Wald statistic or the likelihood ratio test statistic. The likelihood ratio test is used, more generally, to compare the goodness of fit to the observed data of two nested models — a null model and an alternative model with additional parameters.

Mixed-effects models

An extension of linear and logistic regression models that is used commonly in psycholinguistics and was adopted in some more recent analyses of word order variation are linear and logistic mixed-effects models (Jaeger, 2008; Baayen et al., 2008; Quené and Van den Bergh, 2008; Winter, 2013). These models are designed to

take into account the effect of grouping in the data observations. For example, when collecting experimental data of reading times, typically each participant is asked to read a number of sentences. In the resulting data, there will be groups of data points which share the same value of the categorical subject variable. As a result, these N groups of observations (corresponding to N subjects) can have some distinct properties affecting the response variable which cannot be generalised to the whole population. For example, some people tend to read (in general) faster than the others. The reading time measurements obtained for such subjects can be much shorter than for some other subjects, regardless of the other parameters affecting the reading time. To account for this variation at the group level, mixed-effect models separate two types of factors: fixed effects (the ones we hypothesise and test) and random effects (the ones that are introduced by grouping variables sampled randomly from some population). A subject's average reading speed is exactly this kind of random effect since we assume that the subjects in the experiment were chosen randomly and are therefore expected to be normally distributed with respect to the reading speed at the population level.

Intuitively, we can also think of having N different models of reading time — one for each subject i — but where the coefficients for the population-level factors must be shared. The following notation specifies a linear mixed effect model:

$$\begin{aligned} Y_{1,j} &= \boldsymbol{X}_{1,j}\boldsymbol{\beta} + \boldsymbol{Z}_{1,j}\boldsymbol{b}_1 + \epsilon_{1,j} \\ Y_{2,j} &= \boldsymbol{X}_{2,j}\boldsymbol{\beta} + \boldsymbol{Z}_{2,j}\boldsymbol{b}_2 + \epsilon_{2,j} \\ &\dots \\ Y_{N,j} &= \boldsymbol{X}_{N,j}\boldsymbol{\beta} + \boldsymbol{Z}_{1,j}\boldsymbol{b}_N + \epsilon_{N,j} \end{aligned} \tag{4.9}$$

The coefficients $\boldsymbol{\beta}$ of the fixed effects $\boldsymbol{X}$ are the same for all observations $Y_{i,j}$ where i is the index of the group (e.g., subject id) and j is the index of the point coming from the group i (j varies from 1 to n_i where n_i is the number of observations for the group i). The coefficients $\boldsymbol{b}_i$ are, on the other hand, defined for each group i separately. Furthermore, since they correspond to random effects, the coefficients $\boldsymbol{b}_i$ are assumed to be independent and normally distributed: $\boldsymbol{b}_i \sim N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a covariance matrix. Random errors $\epsilon_{i,j}$ are assumed to be sampled randomly

from a normal distribution (the same for all groups), as previously for simple linear regressions. The parameters of the model which are estimated from the observed data are $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and σ.

The logit mixed-effects model is formulated similarly to the linear mixed-effects model in the Equation (4.9) by substituting the values of Y on the left by the logit transformations of $P(Y = 1)$ as in Equation (4.8) (Jaeger, 2008).

In this work, we use the freely accessible, widely used `lme4` package in R (Bates et al., 2014) to estimate the logit mixed-effect models and test the fixed effects $\boldsymbol{\beta}$. The package employs numeric algorithms for the maximum likelihood estimation of the parameters (described in detail in Bates et al. (2014)). To test whether the coefficients are different from 0, the recommended general approach that can be applied to both logistic and linear, simple and mixed effects models is to do the likelihood ratio test on the models with and without the parameter of interest.

4.2 Modelling DLM effects in the adjective placement in complex noun phrases

In this section, we demonstrate how we can apply the general principle of DLM to the noun phrase. Our starting point is a *global DLM principle* inspired by the work of Temperley (2007) and Gildea and Temperley (2010). Two main ideas behind this specific formalisation of the DLM principle are the global minimisation of dependencies in the sentence, as given by the sum of the individual dependencies, and the reliance on the dependency annotation provided by a treebank for decisions related to the syntactic structure of the sentences.

We will show below that the DLM principle for word order variation defined as a preference for a word order which minimises the total sum of all dependencies in the sentence is compatible with the evidence for DLM, such as the heavy-NP shift in English, analysed previously in terms of relative lengths of constituents (Wasow, 2002) or the processing domain minimisation (Hawkins, 1994). For the adjective-noun variation, where there are several dependencies that can be minimised, it is reasonable

to take the global minimisation principle as an initial hypothesis. As in Chapter 3, the length of a dependency is computed in terms of number of words.

We rely on the dependency structure annotation given in a treebank to formalise the computation of the dependency lengths. In other words, we only consider the dependencies which are given in a pre-existing linguistic analysis. We adopt this approach to avoid making any ad hoc assumptions about the syntactic structure jointly with our minimisation analysis. This is an important point for our work distinguishing it from the previous accounts of DLM for specific word order variation constructions: we explicitly separate the investigation of DLM effects from the analysis of the syntactic structure. We rely on the dependency grammar analysis and not on the other previously used accounts (e.g., Hawkins (1994)) for several reasons. As we have seen, there exist a harmonised cross-linguistic dependency grammar for many languages and corpora necessary for a large-scale automatic analysis of word order variation. On the other hand, the syntactic analyses used by Hawkins (1994) are done manually on samples of small scale and for a selected set of constructions. Moreover, the dependency analysis is a very general syntactic analysis which makes feasible the transfer of DLM generalisations between different constructions (e.g., the variation in postverbal domain in English and adjective-noun variation in Romance).

In comparison, the previous studies have used various ways to choose what dependencies are relevant for their analyses and also how to compute their lengths. In Hawkins (1994), the relevant dependencies have been assumed to hold between constituents instead of head words (as specified by a dependency grammar); the left edge (the starting word) of the constituent is taken to effectively define the length of the "dependency" between, for example a verb and its complement. Although, Hawkins' subsequent work assumes that semantic dependencies between the verb and the head noun are also minimised (Hawkins, 2004; Lohse et al., 2004). For Gibson (1998, 2000), alternatively, the relevant processing DLM measure is the distance between the head words of syntactic constituents. For reasons related to memory storage and activation assumptions, the distance is computed in terms of the number of intervening discourse referents, instead of the number of words. Given these differences between various flavors of DLM, the formalisation based on the dependency annotation gives a more general approach to the problem. It allows unifying the

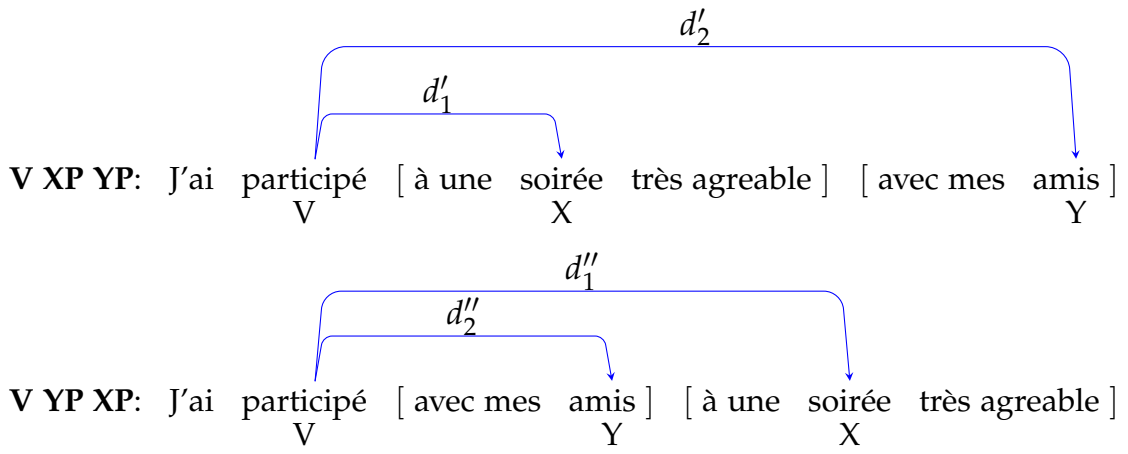


Figure 4.2: Illustration of the alternation of postverbal dependents in French in the sentence 'I participated [$_{XP}$ to a very enjoyable evening] [$_{YP}$ with my friends] '.

previous DLM accounts and providing an analysis of DLM effects in a systematic way under the same assumptions about syntactic structure.

Note that the above differences in the definition and operationalisation of dependency lengths have not been problematic for establishing the previous evidence for DLM. One reason is that, for the most commonly studied constructions, such as the alternation of postverbal complements, slightly different principles still point to the same preferences. In fact, on the basis of the alternation of postverbal dependents in head-initial languages only (without, e.g., taking into account the evidence from verb-final languages such as Japanese), the short-before-long or the principle of end weight (Wasow, 2002), Hawkins' domain minimisation principle (minimising the maximal dependency length, (Hawkins, 1994)) and the global DLM principle (minimising the sum of dependencies) are all equivalent.

This point is illustrated in Figure 4.2 for an example in French. Two prepositional phrase dependents of the verb *participé*, XP and YP, can be exchanged to produce two alternative orders V XP YP or V YP XP. Importantly, the two phrases both occur on the same side of the head which we will indicate using the notation V {XP, YP}. In this case, we can easily show that the short-before-long principle based on the relative length between the two phrases will give the same predictions as a principle based

on the minimisation of dependencies.

The only two dependencies that change their lengths between the two alternative word orders are indicated in Figure 4.2. All other dependencies, including the ones internal to XP and YP, or internal to the VP are not relevant to deciding between these two word orders.[7] If we assume the global DLM principle which prefers the order which minimises the overall *sum* of all dependencies in the sentence then, for our example, the order V XP YP should be preferred if $d'_2 + d'_1 < d''_1 + d''_2$. Since $d'_2 - d''_2$ is equal to the length of the phrase XP and $d''_1 - d'_1$ is equal to the length of the phrase YP, we can rewrite the preference condition as follows: $|XP| < |YP|$. In other words, the order V XP YP is preferred if XP is shorter than YP. The preference for the short-before-long order is exactly equivalent to the preference for the order which minimises the length of dependencies. Note also that this calculation gives the same result if, instead of dependencies V–X and V–Y where X and Y are the content heads of the phrases XP and YP (nouns *soirée* and *amis*), we consider the distance between the verb V and the functional heads of the phrases (prepositions *à* and *avec*) as Hawkins (1994) proposes. The choice between the two possible syntactic structures is, therefore, not crucial for establishing a DLM effect in this construction.

The computations in the example above are particularly straight-forward because the word order alternation is of the type Head {XP, YP}. In the cases where there are two dependents of one head appearing on the same side of the head, the difference between the two orders always amounts to the differences in the lengths of two dependent phrases XP and YP. Not all cases of word order alternation are however of the simple Head {XP, YP} type. The variation in the adjective placement in Romance which we investigate in this chapter is an example of a different structural type of alternation and cannot be reduced to the analysis of only two dependencies. To see the principled difference between the two word order variation constructions, consider an example of the adjective placement in a relatively complex noun phrase (Figure 4.3). When the adjective changes its position relative to the head, this affects a number of dependencies both inside the noun phrase (d_2, d_3) and the external dependency

[7] If there are other dependents in the VP, it is possible that neither of the two word orders will be optimal from the DLM perspective. However, the relative preference between the two word orders considered here will not change.

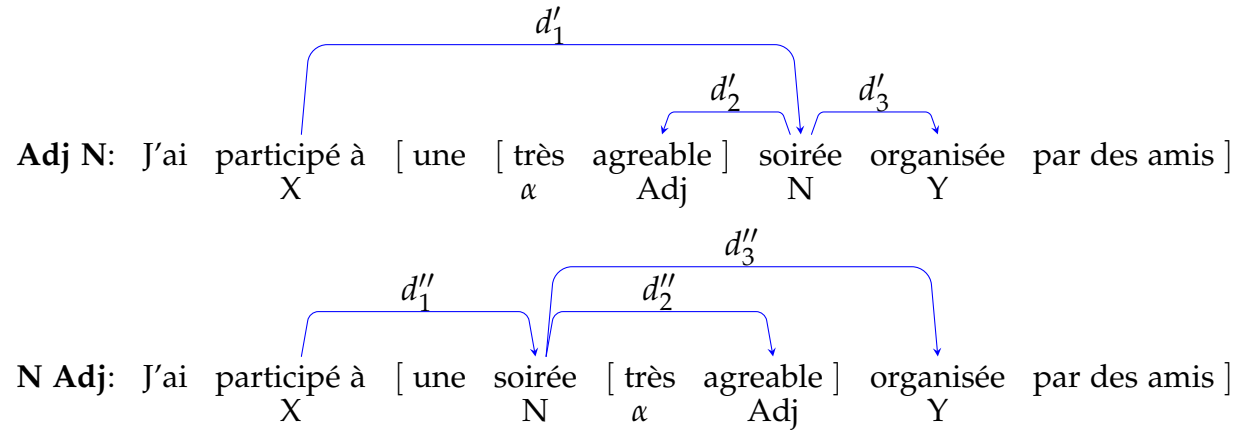


Figure 4.3: Illustration of the adjective-noun order alternation in French in the sentence 'I participated [[to a very enjoyable] evening [organised by my friends] '.

between the noun and its head (d_1). Note, for example, that the dependency X–N is shorter for the postnominal order ($d''_1 < d'_1$) but the dependency N–Y is longer ($d''_3 > d'_3$). These two dependencies have distinct functional types and it is not clear whether it makes sense to compare their lengths. In the general case, the dependency lengths will depend on the composition of the adjective phrase, on the composition of the noun phrase (whether the noun has any additional modifiers), on the position of the noun with respect to its head. Overall, the case of adjective variation in complex noun phrases is unlike the case of postverbal dependents (Figure 4.2) where the only two parameters were the lengths of two dependencies of the same functional type.

The fact that in the adjective-noun order variation many dissimilar dependencies change their length poses two general questions for the formulation of the DLM principle which were not relevant for the case of Head {XP, YP} alternation.

1. First, which dependencies should we consider if we want to study the position of the adjective? We can apply a *global* interpretation of DLM and take into account all dependencies or, alternatively, we can consider a *local* DLM principle and retain as relevant only the dependency between the adjective and its head noun.

2. Secondly, if all modified dependencies influence the adjective placement, do

they do it in the same way? In other words, can the global DLM principle be expressed in terms of a comparison of the *sum* of all dependencies for the two sentences?

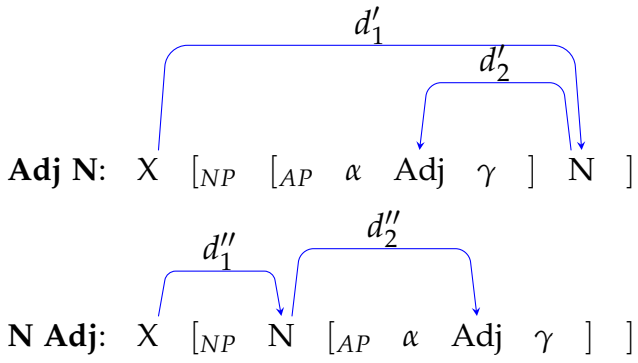
In relation to the previous work on adjective variation discussed in Section 4.1.1, we can pose one more question. Is the treatment of the adjective-noun variation in terms of DLM more appropriate than the adoption of a principle such as heavy-adjective shift? (Or are they equivalent as it is the case for heavy-XP shift and DLM in postverbal alternation in English?). Note that the adjective phrase can have some pre-adjectival modifiers (α, adverb *très*) and post-adjectival modifiers (we will refer to them as γ, none are present in the example in Figure 4.3). The presence of α and γ elements in the adjective phrase affect the lengths of dependencies in different ways. On the other hand, the heavy-adjective principle would treat them holistically.

In this chapter, we analyse the corpus-based distribution of adjectives in Romance languages with the goal to shed light on these fundamental and intriguing questions for the research on dependency length minimisation. We start by formulating the global DLM principle for the adjective-noun construction using the sum of all dependency lengths as our measure. We derive the prediction for prenominal and postnominal placement preferences using this formalisation and compare them to the predictions arising when each dependency is minimised individually.

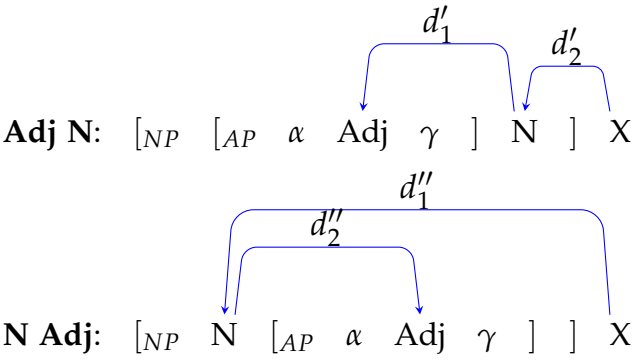
Throughout the formalisation and subsequent analyses we use the definition of dependency length (DL) in terms of number of words, as in Chapter 3. Recall that if two words w_i and w_j ($j > i$) are connected by a dependency in a dependency tree annotation of the sentence $w_1, \ldots w_l$, the length of this dependency equals the difference between their indices: $DL = j - i$. If two words are adjacent in the sentence, their dependency length will be therefore equal to 1.

4.2.1 Formalisation and predictions of the DLM principle

Consider, as a first prototypical case, a simple noun phrase with only one adjective phrase as a modifier (Figure 4.4).



(a) Left external dependency



(b) Right external dependency

Figure 4.4: Prenominal and postnominal variants of a simple noun phrase given left (a) and right (b) external dependency X–N.

The adjectival modifier can be a complex phrase with some left dependents and some right dependents. A left dependent would be typically an adverb, e.g., *very proud* and a right dependent would be typically a complement, e.g. *proud of his achievements*. We indicate all of the left and all of the right dependents in our schematic representation by α and γ, respectively. To simplify the notation, we will also use α and γ to indicate the *overall lengths* of the left and right dependents when we compute the dependency lengths.

Note that in Figure 4.4 and in all the following examples, the dependency annotation of the schematic constructions follows the content-head annotation of the UD scheme (Chapter 2) which we adopt consistently as our underlying syntactic representation. This implies that the noun N is the head of the noun phrase (and not the determiner,

	DL pre-N	DL post-N	Preference
N–Adj	$\gamma + 1$	$\alpha + 1$	pre-N if $\alpha > 0$, post-N if $\gamma > 0$
X–N, X on the left	$\alpha + \gamma + 2$	1	post-N, more so if $\alpha, \gamma > 0$
X–N, X on the right	1	$\alpha + \gamma + 2$	pre-N, more so if $\alpha, \gamma > 0$

Table 4.1: The lengths of N–Adj and X–N dependencies in the case of prenominal (pre-N) and postnominal (post-N) placement of the adjective. For each dependency, we specify what order would be preferred if this dependency tends to be minimised (independently from other dependencies).

for example) and that the relevant dependencies are all between the noun and its modifiers and the noun and its head.

The head of the noun N is indicated by X. The left position of X can correspond, for example, to a verb-object relation between X and N. The sentence *I knew*$_X$ *a man*$_N$ *very proud of his achievements* is an example of this structural variant represented in Figure 4.4a. X can also frequently be a noun taking N as its prepositional complement as in *son*$_X$ *of a man*$_N$ (since the head of a prepositional phrase is always a noun in the content-head annotation). The right position of X corresponds typically to a subject-verb relation between X and N. The sentence *A man*$_N$ *very proud of his achievements would not do*$_X$ *this* belongs to this second structural variant depicted in Figure 4.4b. In French and other Romance languages, the sentences analogous to the English examples above can in principle appear with prenominal or postnominal adjective (as illustrated in the example in Figure 4.3).

The only two dependencies that change their lengths between Adj N and N Adj orders and are therefore relevant for the DLM principle are the noun-adjective dependency and the X–noun (X–N) dependency. Table 4.1 indicates the lengths of these two dependencies in four cases illustrated in Figure 4.4. In particular, the adjective-noun dependency length depends on both the length of the pre-adjectival material α and post-adjectival material γ of the adjective phrase. The length of the X–N dependency depends on the adjective position and on the *overall length* of the adjective phrase equal to $\alpha + \gamma + 1$.

	ΔDL	Preference
X=Left	$2\gamma + 1$	post-N, more so if $\gamma > 0$
X=Right	$-2\alpha - 1$	pre-N, more so if $\alpha > 0$

Table 4.2: Dependency length difference and the corresponding preference for a simple type of the noun phrase ($\Delta DL = DL_1 - DL_2$).

Table 4.1 also lists the preferences between prenominal and postnominal orders for each of the dependencies if they are to be minimised independently. In particular, whatever the position of X is, the minimisation of the N–Adj dependency should favor the postnominal placement of AdjP in the presence of post-adjectival elements γ and should favor the prenominal placement of AdjP in the presence of the pre-adjectival elements α.

The two alternative linearisations for each of the two structural variants (a) and (b) in Figure 4.4 also yield different *total* dependency lengths. By convention, we will always indicate the total dependency lengths (the sum of all dependencies) in the prenominal order as DL_1, and in the postnominal order as DL_2. Their difference is always calculated as $\Delta DL = DL_1 - DL_2$.[8] If $\Delta DL > 0$ then the postnominal order is preferred, otherwise if $\Delta DL < 0$ the prenominal order is preferred. Given the dependency lengths in Table 4.1, the resulting differences in the sum of DLs are summarised in Table 4.2. Qualitatively, the differences in preferences between the two tables arise in two cases: when X=Left and $\alpha > 0$ and when X=Right and $\gamma > 0$. In these cases, the two dependencies point in the opposite directions for minimisation. The sum measure obtains a compromise which results in a general post-N preference when X=Left (contrary to the N–Adj dependency individual pre-N preference when $\alpha > 0$) and a general pre-N preference when X=Right (contrary to the N–Adj individual post-N preference when $\gamma > 0$).

The predictions by the DLM analysis carried out above are influenced crucially by

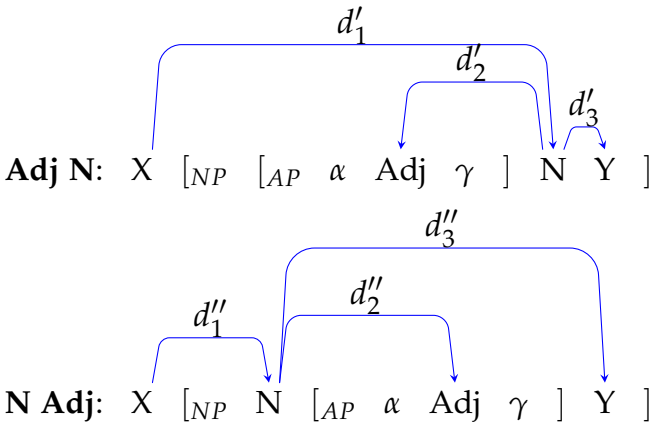
[8] We do not need to take into account the dependencies which have the same length for both word orders when computing the difference $DL_1 - DL_2$ because they will cancel out. We therefore will compute only the sum of dependency lengths of the relevant dependencies (e.g. d'_1, d'_2) to be DL_1 and DL_2.

the position of the noun phrase with respect to its parent X. This is a surprising prediction which has not been evoked often in the literature. Nevertheless, a related observation comes from Forsgren (1978) who has proposed that the syntactic function of the noun (subject or object) affects the adjective position, with subject noun phrases favoring prenominal placement and object noun phrases favoring, instead, postnominal placement. Given that the subject function corresponds to the position X=Right and the object function corresponds to the position X=Left (where X is a verb), his observation would coincide with the DLM predictions as stated in Table 4.2. On the other hand, Thuilier (2012) in her corpus-based analysis of French has not found the subject/object function to be a significant predictor of the variation.

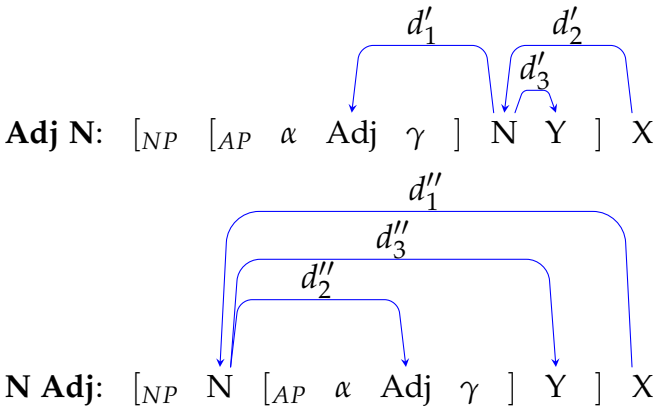
The predictions of the DLM account for complex adjective phrases are only partially aligned with the previous literature. Overall, the presence of post-adjectival material favors postnominal placement both from the point of view of minimisation of the adjective-noun dependency and the total dependency length. This corresponds to the observed tendency for adjectives with complements to appear on the right of the noun (Head-Final Filter, Section 4.1.1). However, the presence of pre-adjectival dependents such as adverbs is predicted to favor prenominal placement if the adjective-noun dependency or the total dependency length are minimised. This is against the observation that adjectives with adverbs tend to appear postnominally in French (Abeillé and Godard, 2000; Thuilier, 2012).

Noun phrases with additional dependents

The noun can also have other modifiers or dependents apart from the adjective. In our illustration in Figure 4.5, we consider a noun with an additional right dependent indicated by Y. These types of noun phrases are very common in Romance languages (almost 50% of noun phrases in our sample include at least one post-head dependent). Importantly, the position of the phrase YP is fixed with respect to the noun. We consider the non-adjectival modifiers of a noun that appear categorically on its right, such as prepositional phrases (*soirée [dans un parc$_Y$]*) or relative clauses (*soirée [organisée$_Y$ par ...]* in the example in Figure 4.3 corresponding to the structure (a) in



(a) Left external dependency



(b) Right external dependency

Figure 4.5: Noun phrase structure variants with an additional right dependent Y.

Figure 4.5). We exclude from our analysis other modifiers of the noun which appear prenominally such as demonstratives or numerals.

Figure 4.5 illustrates two possible placements for an adjective phrase: prenominal Adj N Y and postnominal adjacent to the noun N Adj Y. In fact, a third order is also possible in Romance languages with adjective placed postnominally after Y: N Y Adj. To simplify for the moment, we will assume only the first two possible orders for our illustrations.

In this noun phrase structure, there are now three dependencies which should be taken into account in a DLM analysis. In addition to X–N and N–Adj, the dependency

	DL pre-N	DL post-N	Preference
N–Adj	γ	α	pre-N if $\alpha > 0$, post-N if $\gamma > 0$
N–Y	1	$\alpha + \gamma + 2$	pre-N, more so if $\alpha, \gamma > 0$
X–N, X on the left	$\alpha + \gamma + 2$	1	post-N, more so if $\alpha, \gamma > 0$
X–N, X on the right	$\lvert Y \rvert + 1$	$\alpha + \gamma + 2 + \lvert Y \rvert$	pre-N, more so if $\alpha, \gamma > 0$

Table 4.3: The lengths of N–Adj, N–Y and X–N dependencies in the case of prenominal (pre-N) and postnominal (post-N) placement of the adjective. For each dependency, we specify what order would be preferred if this dependency tends to be minimised (independently from other dependencies).

	ΔDL	Preference
X=Left	$\gamma - \alpha$	post-N, if $\gamma > 0$, pre-N if $\alpha > 0$
X=Right	$-3\alpha - \gamma - 2$	pre-N, more so if $\alpha, \gamma > 0$

Table 4.4: Dependency length difference and the corresponding word order preference in complex noun phrases with a right dependent Y ($\Delta DL = DL_1 - DL_2$).

N–Y also gets two different lengths in the two possible orders. The computed lengths DL_1 and DL_2 for the three dependencies are shown in Table 4.3.[9] The values for the dependency Adj–N are the same as for the noun phrases without a right dependent (Table 4.1); for the dependency X–N there is an additional component $|Y|$ (the length of the phrase YP) but it is cancelled out when we look at the difference between prenominal and postnominal DLs. The resulting preferences towards an order minimising Adj–N and X–N dependencies individually are therefore exactly the same as for the case of simple noun phrases without Y. The N–Y dependency shows, in turn, a clear preference for the prenominal placement of the adjective which would allow adjacency between N and YP.

The total length of the three dependencies for the Adj N and N Adj orders is given

[9]Note that we do not need to take into account here the position of the head Y in the phrase YP. Similarly to the case of postverbal dependents, this position is irrelevant for the computation of the difference of the lengths of N–Y dependency for the two orders since its contribution will be cancelled out.

in Table 4.4. As previously, we distinguish between the structure with X on the left of the noun N and with X on the right of N. These values are not exactly equal to the ones obtained for the simple noun phrase (Table 4.2) because they include the contribution of the N–Y dependency. Overall, there is a shift towards more prenominal preference which is a consequence of the strong prenominal preference of the N–Y dependency.

Summary of predictions

Given the structures and dependencies analysed above, we can formulate the predictions for the adjective placement stemming from the DLM principle. We test these predictions empirically using the corpus data of five Romance languages as described in the next section.

If we adopt the most general global DLM principle that the total sum of all dependencies should be minimised then the preferred order should be correlated with the value of ΔDL as summarised in Tables 4.2 and 4.4.

It is possible that the global DLM principle is not sensitive enough because there are some dependencies in the noun phrase whose minimisation can favor the opposite orders. Ideally, we would like to measure and test the DLM effect for each of these dependencies separately. This is, however, not straightforward. The lengths of the three dependencies are built up from the same two parameters α and γ and the corresponding factors will be, as a result, correlated and not independent.

Instead, we propose to test a simplified set of factors abstracting away from the interaction of X–N and N–Y dependencies with the composition of the adjective phrase. Given the preferences in Tables 4.1 and 4.3, we simplify them into the following predictions which capture the *overall* tendency in minimisation of individual dependencies:

Dependency	Preference
X–N	prenominal if X=Right, postnominal if X=Left
N–Y	prenominal
N–Adj	prenominal if $\alpha > 0$, postnominal if $\gamma > 0$

The factors for the dependencies X–N and N–Y are therefore a binary approximation of the factors corresponding to individual dependency lengths. The N–Y factor should indicate, for example, whether there is an overall tendency towards more prenominal order when Y is present, while the X–N factor will tell us whether there is a difference in the distribution of adjectives when X is on the right or on the left of the noun.

In addition to these main predictions, we will refer back to the more fine-grained predictions devised previously when we analyse the results, obtained from the corpus data, in connection with the interaction of the dependency factors.

4.2.2 Experimental setup

In this section, we describe the corpus data and the details of statistical methods we use to analyse the distribution of adjectives in Romance languages.

Data extraction

Our analyses are based on the data from five Romance languages: Italian, French, Catalan, Spanish and Portuguese. We use the Universal Dependencies treebanks v1.3 for these languages which are available freely online (Nivre et al., 2016).[10] We do not include other Romance languages into our analysis chiefly for lack of availability of sufficiently large syntactically annotated corpora for those languages. For example, there exist a UD treebank of Romanian, but its size (in the version 1.3 of UD treebanks that we have obtained for our experiments) is relatively small (less

[10] `http://universaldependencies.org`

than 5000 sentences). Since we study complex noun phrases and the word order variation associated with many parameters, we focused on five languages that had the largest treebanks available. Each UD treebank comes divided into three parts (train, development and test) which are habitually used to train and evaluate NLP parsing systems. We use only the training sections of the treebanks in our experiments which constitute between 80% and 90% of the overall annotated data.

Using the dependency annotation, we extracted all noun phrases which contain an adjective. The simple automatic procedure finds the nouns and adjectives using the part-of-speech tags ('ADJ' for adjectives and 'NOUN' for nouns) and extracts only the cases when the adjective is the child of a noun (based on the annotation information described in the Chapter 2). The relevant properties such as the position of the noun head and the presence of a right dependent are also extracted automatically based on the dependency annotation. The resulting data is a table with rows corresponding to the observations and columns corresponding to the variables we want to study. Importantly, we also extract the *lemmas* of adjectives. As we have discussed previously, the adjective placement is strongly conditioned lexically. The lemma variable is therefore one of the defining factors in the analysis.

Some basic preprocessing steps were applied on the extracted data to ensure that all nouns and adjectives are well-formed words (and not, for example, symbols or numeric expressions). Importantly, we also removed all examples of the noun phrases which contain punctuation. Punctuation can indicate a focused or a parenthetical adjective phrase which do not have the same syntactic distribution as the non-stressed standard cases of adjectival modification.

The resulting data contains between 7200 and 15900 observations per language. The statistics of the data in terms of the number of adjective tokens (i.e., the number of observations), the number of adjective types (i.e., the number of distinct lemmas) and the frequency of the prenominal and postnominal orders are given in Table 4.5 (first three columns). Overall, an adjective appears in prenominal position in between 22% (Catalan) and 32% (Italian) of observations confirming empirically that the postnominal position is the default one in Romance languages.

Since the variation is very constrained for some types of adjectives (e.g., adjectives

Order	All Adj #	All Adj %	Alternating Adj #	Alternating Adj %
Spanish				
Adj N	4533	28.5	3835	57.2
N Adj	10400	65.4	2586	38.6
N YP Adj	970	6.1	282	4.2
Total tokens	15903		6703	42.1
Total types	3410		334	9.8
Catalan				
Adj N	3641	21.8	3276	54.7
N Adj	11711	70.2	2405	40.2
N YP Adj	1319	7.9	306	5.1
Total tokens	16671		5987	35.9
Total types	2637		250	9.5
French				
Adj N	4202	28.5	3520	70.4
N Adj	9380	63.7	1299	26.0
N YP Adj	1138	7.7	179	3.6
Total tokens	14720		4998	34.0
Total types	2777		181	6.5
Italian				
Adj N	3669	32.0	2991	55.7
N Adj	7251	63.3	2192	40.8
N YP Adj	541	4.7	186	3.5
Total tokens	11461		5369	46.8
Total types	2151		329	15.3
Portuguese				
Adj N	2142	29.8	1870	67.8
N Adj	4729	65.9	806	29.2
N YP Adj	309	4.3	84	3.0
Total tokens	7180		2760	38.4
Total types	1686		162	9.6

Table 4.5: Token and type frequencies of adjectives and their placement in the extracted data for five Romance languages.

of color or nationality appear almost exclusively postnominally) it makes sense to exclude them from our analysis which assumes that both prenominal and postnominal orders are in principle available. Out of all adjectives, we extracted therefore only the ones which appear at least once in a prenominal position and at least once in a postnominal position in our corpus. Table 4.5 (columns 4, 5) gives the size of this subset of alternating adjectives and the overall number of observations in which they occur. We retain, as can be seen from these numbers, between 6.5 and 15% of distinct adjectives. Note that since most of the adjectives appear only once in the corpora (this is the familiar Zipf's law for lexical frequency), this relatively small number of adjective types accounts for around 34 to 47% of total observations. For this reason, the focus on the data with alternating adjectives is also more appropriate from the practical point of view. Since we will use adjective lemma as a parameter in our model, it is desirable to have more than one observation for each parameter value to obtain robust numerical estimations.[11]

Note that the statistics on prenominal versus postnominal frequency shifts towards more prenominal order when only alternating adjectives are taken into account (Table 4.5 indicates between 55 and 70% of prenominal orders). This is not surprising since many alternating and often prenominal adjectives such as *good, beautiful, small* etc are some of the most frequent adjectives.

Statistical analysis using logit mixed effect models

To analyse the variation we use logit mixed effect models described in Section 4.1.2. Our statistical analysis is similar to the ones conducted in Bresnan et al. (2007); Thuilier (2012); Fox and Thuilier (2012).

The output variable which our statistical models are designed to predict is always `Order`. In the experiments in this section we only consider two possible orders: Adj N and N Adj (YP). As can be seen from Table 4.5, the number of observations which have the third possible order N YP Adj is very small (only 3–5% of cases). When we

[11] In fact, the mixed effect models which we describe below did not always converge when used with, e.g., 3000 random effects (adjective lemmas) for 15000 observations.

combined the two categories N Adj and N Adj YP under one category the results of our models did not change significantly compared to the results obtained when N Adj YP cases were removed from the data. We assume, therefore, that we can merge the two postnominal orders under one category N Adj without significant loss of generality. This allows us to apply models with the binary response variable which are easy to interpret and can be compared to the models previously used in word order variation studies.

We encode the values of our binary `Order` variable as follows: `Order` = 0 corresponds to Adj N and `Order` = 1 to N Adj order. When interpreting the results of our model, if the coefficient β_i is positive this will mean that the factor i 'votes' for the value 1 favoring, therefore, the postnominal order. If β_i is negative then the factor i favors, instead, the prenominal order. The sign of β_0, or intercept, of the model tells us whether the adjective position in the data is more prenominal or postnominal on average.

The predictor variables corresponding to individual dependency lengths effects that we test are: `PositionX, PresenceY, Alpha, Gamma`. `PositionX, PresenceY` are binary variables; `PositionX` codes two values: `X=Left` and `X=Right` and `PresenceY` codes `True` and `False` values. `Alpha, Gamma` are numeric variables with discrete values $0, 1, 2, \ldots$ which count the sizes of left and right dependents of the adjective as number of words.

An alternative model which stems from the global DLM principle has only one parameter ΔDL which is computed as $DL_1 - DL_2$ for each noun phrase by constructing a prenominal (DL_1) and a postnominal placement (DL_2) of the adjective phrase. ΔDL is a numeric variable with integer values; the increase in ΔDL is predicted to correlate with postnominal placement ($\beta > 0$).

The main type of the model we use to test the effect of these predictor variables on the variable `Order` is logit mixed effects model with adjective lemmas as random effects. As we have discussed above, distinct adjective types have different preferences for prenominal and postnominal orders. Consider a toy corpus where adj_1 appears 100 times, out of which 20 times in prenominal position, and adj_2 appears 100 times, out of which 80 times in prenominal position. The average distribution of

prenominal and postnominal orders is 50% in this corpus, however, this number does not accurately describe the distribution of either adj_1 or adj_2. To account for this variation systematically, we consider the observations for each adjective type to form a group, similarly to the grouping by subject in psycholinguistic experiments. A model with adjective lemmas as random effects can be seen therefore as a lexicalised model of variation, i.e., where the order depends on the word type and not only on more abstract syntactic information.

For brevity, we will use the `lme4` notation to describe our models. The following expression specifies a model with fixed effects X_1, X_2, X_3 and random effects Adj.

$$Order \sim X_1 + X_2 + X_3 + (1 \mid Adj) \tag{4.10}$$

This is a simplified notation for a model with an intercept β_0, slope coefficients $\beta_1, \beta_2, \beta_3$ of the effects X_1, X_2, X_3 and random intercepts b_{0i} for each adjective lemma Adj_i. We do not include random slopes b_{1i}, b_{2i}, b_{3i} into our model because we do not have sufficient data for robust estimation of the large number of parameters. The omission of random slopes corresponds to a reasonable assumption in the absence of relevant evidence that the lexical type of adjective does not affect the *strength* of the effects X_1, X_2, X_3.

The null model for the testing of individual parameters is a model without any fixed effects but with the random adjective effects:

$$H_0 : Order \sim (1 \mid Adj) \tag{4.11}$$

The results of fitting this model on our five datasets are given in Table 4.6.[12] For each language, the table lists a standard summary for a mixed effect model fit. The parameters of each model are aligned column-wise for easier comparison of results across languages. First, the table lists the intercepts β_0 of the models which are not significantly different from 0, apart for French. This just means that the prenominal placement observed from the numbers in Table 4.5 can be largely explained by a small number of frequent adjectives favoring prenominal placement. The value

[12]This and subsequent tables are generated automatically from `lme4` models in R using `texreg` package Leifeld (2013).

	Italian	Spanish	Catalan	French	Portuguese
(Intercept)	−0.03	−0.03	0.01	−0.34**	−0.17
	(0.08)	(0.08)	(0.10)	(0.13)	(0.12)
Log Likelihood	-2733.02	-3606.76	-2752.28	-1815.71	-1242.55
Num. obs.	5369	7739	5987	4998	2760
Num. groups: Adj	329	414	250	181	162
Var: Adj (Intercept)	1.42	1.52	1.59	2.37	1.46

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4.6: The null model: $OrderBinary \sim (1 \mid Adj)$. The values in parentheses indicate the standard errors of the estimates of a parameter (here, the intercept).

in parentheses below each intercept value is the standard error for the parameter. This table and subsequent tables summarizing the fit of our model also include the important parameters such as the total number of observations, the number of groups (distinct adjective lemmas) and the variance of adjective intercepts. Other values in the table such as the log-likelihood values are not supposed to be compared across models and languages since they are evaluated on different datasets. Instead, they will be compared to the log-likelihood values of models with additional fixed effects.

4.2.3 Results and discussion

We start first by presenting the results for the global DLM model and the most powerful model — the model that includes all individual factors deduced from our DLM analysis. We then present the analyses of each factor separately for a detailed understanding and interpretation of the results.

	Italian	Spanish	Catalan	French	Portuguese
(Intercept)	-0.03	-0.03	0.01	-0.33^{*}	-0.16
	(0.08)	(0.08)	(0.10)	(0.13)	(0.12)
ΔDL	0.06^{***}	0.05^{***}	0.02	0.04	0.07^{***}
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Log Likelihood	-2721.42	-3594.54	-2751.44	-1814.31	-1235.79
Num. obs.	5369	7739	5987	4998	2760
Num. groups: Adj	329	414	250	181	162
Var: Adj (Intercept)	1.40	1.52	1.59	2.36	1.47

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4.7: The fit of the global DLM model: $Order \sim \Delta DL + (1 \mid Adj)$.

Global DLM model

The model which tests the global cumulative DLM principle has only one independent variable ΔDL which subsumes all the relative lengths of the dependencies in the noun phrase. The model fit on our data is given in Table 4.7.

Despite the fact that ΔDL has positive effect for all five languages, it is a small effect which reaches significance only for Italian, Spanish and Portuguese. This result suggests that adjective variation in Romance languages respects only to some extent the principle of minimisation of total dependency length in a sentence. There could be two main reasons for this result: the DLM does not readily apply for the case of adjective variation, or that the DLM effects of the dependencies do not interact in a simple way as assumed by the global model. In fact, the global model does not tell us which dependencies are minimised since their individual effects are confounded together. The models we present next are designed to investigate and test the DLM effects individually.

Model with all individual parameters

The most potentially powerful model we test is the model which includes the four parameters `Alpha`, `Gamma`, `PositionX`, `PresenceY` capturing the variation in the lengths of three dependencies N–Adj, N–Y and X–N which we identified in the previous section. In principle, a more powerful model would include also the interactions between these four parameters. However, such model has many more degrees of freedom and the algorithm for estimation of these parameters could not converge on our data. We present therefore the fit of the model only with the linear combination of the four parameters (Table 4.8).

First of all, note that this model improves considerably the log-likelihood of the data compared to the previous global DLM model, including for Italian, Spanish and Portuguese where the DLM effect was significant. The testing of statistical significance of these values between these two models is problematic since they are not nested. We can compare instead the values of Akaike Information Criterion (AIC) which take into account the log-likelihood and the number of parameters in the model. Note that our second model has three more parameters and is theoretically more powerful than the first one. Despite this, the values of AIC indicate clearly that the second model generalises better the data than the first one (smaller AIC values indicate better fit, *f. e.* stands for *fixed effects*):

AIC	Italian	Spanish	Catalan	French	Portuguese
Global model (1 f. e.)	5448.84	7195.09	5508.88	3634.62	2477.58
Complete model (4 f. e.)	5190.02	6792.85	5217.24	3523.97	2334.14
Δ AIC	258.82	403.24	291.64	110.65	143.44

We can conclude, therefore, that adjective-noun variation exhibits DLM effects but the various dependencies are not minimised to the same extent as assumed by the simple global DLM model. We turn now to the interpretation of the values of coefficients estimated for our four parameters in Table 4.8.

	Italian	Spanish	Catalan	French	Portuguese
(Intercept)	0.13	0.33^{**}	0.31^{*}	-0.13	-0.02
	(0.13)	(0.12)	(0.14)	(0.17)	(0.19)
Alpha	0.89^{***}	0.56^{***}	0.85^{***}	0.52^{***}	1.02^{***}
	(0.12)	(0.07)	(0.10)	(0.10)	(0.16)
Gamma	0.10^{***}	0.06^{***}	0.01	0.06^{*}	0.10^{**}
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
PresenceY=True	-0.96^{***}	-1.16^{***}	-1.05^{***}	-0.85^{***}	-1.02^{***}
	(0.08)	(0.07)	(0.08)	(0.10)	(0.12)
PositionX=Left	0.16	0.09	0.13	0.07	0.21
	(0.10)	(0.09)	(0.10)	(0.11)	(0.15)
Log Likelihood	-2589.01	-3390.43	-2602.62	-1755.98	-1161.07
Num. obs.	5369	7739	5987	4998	2760
Num. groups: Adj	329	414	250	181	162
Var: Adj (Intercept)	1.54	1.66	1.74	2.43	1.65

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4.8: The model with all individual parameters specified as: $Order \sim \alpha + \gamma + PresenceY + PositionX + (1 \mid Adj)$.

Composition of the adjective phrase

Contrary to the predictions in Section 4.2.1 (page 130) for the noun-adjective dependency, the factor `Alpha` does not have the prenominal effect, instead, it has a strong postnominal effect $\beta \in [0.52, 1.02]$, statistically significant for all languages ($p < 0.001$). As we anticipated, this result is in line with the previous corpus-based observations for French (Thuilier, 2012) and extends them for other Romance languages. The other factor related to the composition of the adjective phrase, `Gamma`, favors postnominal placement as predicted both by the minimisation of the adjective-noun dependency and the previous work describing the heavy adjective shift. Note that the effect is rather small (and not significant for Catalan) which can be due to the fact that there is an interaction between `Alpha` and `Gamma` parameters and other dependencies N–Y and X–N.

In fact, if we consider a subset of the data where there are no right dependents of the noun (`PresenceY=False`, around 50% of the overall data), the effect of the `Gamma` factor increases substantially reaching values in the range 0.72–0.99 (Table 4.9[13]). These results for `Gamma` are now in agreement with the predictions for simple noun phrases (Table 4.2, page 133). The change in the values of `Alpha` and `Gamma` show that there is an interaction between these parameters and the presence of a right dependent which is also expected if N–Y dependency is minimised. Note, however, that this situation is also consistent with a principle of heavy adjective postposition sensitive to the overall size of the adjective phrase `Alpha + Gamma` plus the minimisation of the N–Y dependency which depends too on the total adjective phrase length.

To verify whether the factors `Alpha` and `Gamma` can be subsumed under one variable `LengthAP` without loss of generality, we fit the model with this one factor on the same data (only simple noun phrases). As expected, `LengthAP` favors significantly postposition ($p < 0.001$ for all languages). The comparison of AIC values of the two models gives a slight preference for the model with the two parameters:

[13]We do not include the PositionX factor since it was not significant in the overall model.

	Italian	Spanish	Catalan	French	Portuguese
(Intercept)	0.34^{*}	0.62^{***}	0.86^{***}	-0.01	0.43^{*}
	(0.14)	(0.14)	(0.20)	(0.22)	(0.21)
Alpha	2.76^{***}	2.36^{***}	3.57^{***}	1.91^{***}	2.41^{***}
	(0.35)	(0.27)	(0.43)	(0.29)	(0.43)
Gamma	0.99^{***}	0.86^{***}	0.98^{***}	0.99^{***}	0.72^{***}
	(0.13)	(0.12)	(0.18)	(0.17)	(0.13)
Log Likelihood	-1178.63	-1426.42	-927.60	-760.50	-517.08
Num. obs.	2767	3469	2587	2253	1258
Num. groups: Adj	299	377	231	166	141
Var: Adj (Intercept)	3.31	3.71	4.77	5.21	3.15

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4.9: The analysis of the `Alpha` and `Gamma` parameters in the noun phrases without any right dependents: $Order \sim \alpha + \gamma + (1 \mid Adj)$.

AIC	Italian	Spanish	Catalan	French	Portuguese
`Alpha` and `Gamma`	2365	2860	1863	1529	1042
`LengthAP`	2390	2890	1897	1536	1059
Δ AIC	25	30	36	6	16

The differences in AIC values are, nevertheless, rather small to conclude with certainty that there exists any distinct effect of the presence of left versus right dependents of the adjective in terms of dependency length minimisation. To summarise, the overall tendency for adjectival postnominal placement for both `Alpha` and `Gamma` suggests that the adjective-noun dependency is not necessarily minimised, especially when there is an interaction between these factors and the minimisation of the N–Y dependency. The differences in the strength of the effects between `Alpha` and `Gamma` should be most probably attributed to lexical factors, such as a type of adverb, as proposed by Abeillé and Godard (2000). Our syntactic factors are inevitably too coarse to capture this part of the variation.

Presence of a right dependent Y

As can be seen from Table 4.8, the presence of a right dependent Y is a strong predictor for prenominal order in adjective variation ($\beta \in [-1.16, -0.85]$, $p < 0.001$ for all languages). This result confirms clearly that the N–Y dependency is minimised. The preference for more prenominal order in the presence of postnominal dependents has already been observed for French (Forsgren, 1978; Thuilier, 2012). Forsgren (1978) motivates it by the overall "balance" considerations for the organisation of the noun phrase, i.e., if there are two dependents of the noun they will create a more balanced structure if they appear on the two sides of the noun. Importantly, our account of this effect is, instead, deductive and not simply descriptive of the data observed. We have formally predicted the observed pattern as a consequence of a general DLM principle. An additional contribution of our work is that we demonstrate that the N–Y minimisation effects is a general effect observed in the other four Romance languages in addition to French.

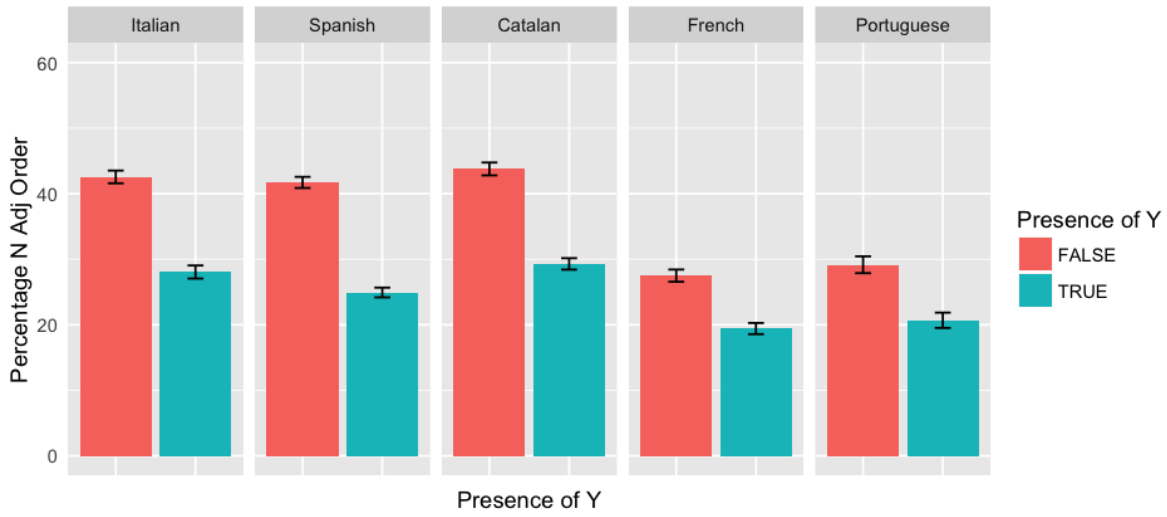


Figure 4.6: The percentage of postnominal order of adjectives in noun phrases with only simple adjectives in two conditions: when there is a right dependent Y (green bars) and when there is no right dependent (red bars).

Interestingly, we find that the effect also persists when the adjective phrase is a simple adjective and, as a result, N–Y dependency can be at most of length 2. The effect can be observed clearly at the level of average adjective placement (without taking into account the individual lexical preferences of the adjectives) in Figure 4.6.[14] For all languages, when there is no right dependent Y (red bars) we observe a higher percentage of postnominal order than when there is a right dependent (green bars). This result is surprising since it suggests that there is a strong minimisation effect for very short dependencies. It is unlikely that this effect can be attributed to processing constraints such as limited short-term memory. To investigate the mechanisms behind the DLM effects in this construction, we conduct a follow-up corpus-based analysis on the distribution of simple adjectives in N + Y noun phrases in Italian presented in Section 4.3.

[14]The effect of `PresenceY` is also statistically significant in the corresponding mixed-effect analysis with adjectives as random effects.

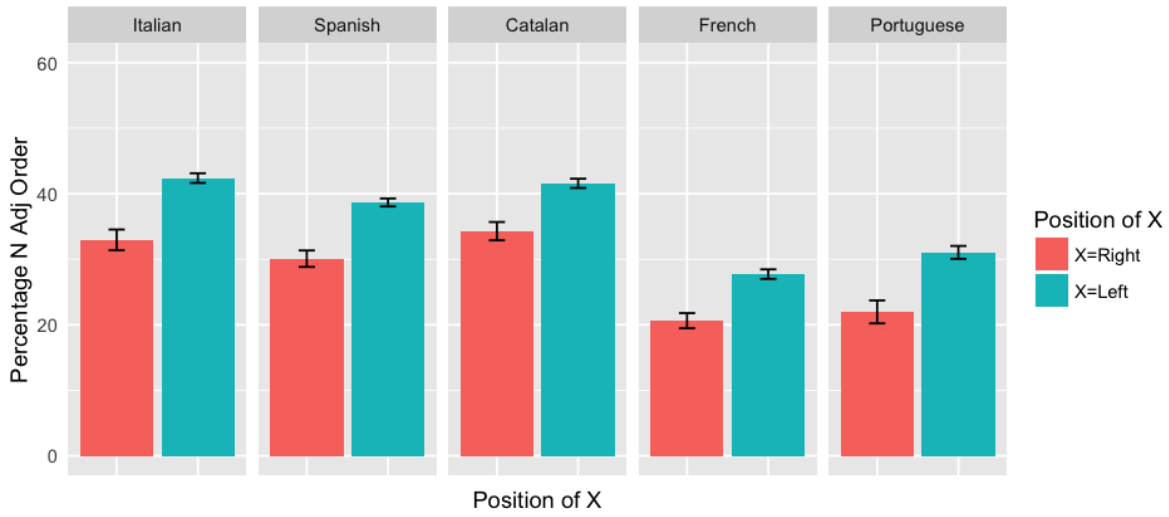


Figure 4.7: The percentage of postnominal adjectives in two conditions: X is on the right of the noun (red bars) and when X is on the left of the noun (green bars).

Position of NP parent X

The estimation of the complete model in Table 4.8 says that while the PositionX=Left factor is estimated as favoring postnominal placement this result is not statistically significant for any language. Despite this, the overall average distribution of adjectives for the two conditions (when X=Right and X=Left) shows the tendency consistent with the DLM predictions (Figure 4.7). For all languages, when X is on the left of N (green bars) we observe a higher percentage of postnominal order than when X is on the right of N (red bars), exactly as predicted by the minimisation of X–N dependency.

One reason why we observe the predicted tendency in the overall distribution but do not get the significant effect for the X–N factor in our statistical model could be the interaction between this and the other factors. As we have seen, both `Alpha` and `Gamma` push towards more postnominal order. In case of X=Left, this effect corresponds exactly to the minimisation of the X–N dependency. In case of X=Right, there is a strong prenominal effect of the N–Y dependency which can overlap with the prenominal effect from the X–N factor. An additional issue is that the X=Right

cases are much less frequent than X=Left cases (20% vs 80% of cases). To see whether there is an independent effect of X–N dependency we can analyse the subset of noun phrases without any right dependents and only containing simple adjectives. Figure 4.8 is very similar to Figure 4.7 and indicates that the previously observed effect is equally present when only one factor — X–N dependency — is active. Note that the effect of X–N binary factor is significant in a simple logistic regression (without random effects).[15] However, similarly to the results in Table 4.8, the same effect is no longer significant when random adjective effects are taken into account. The difference in the results between the models with and without random effects can arise from the skewed distribution of adjectival types between the two conditions (X=Left and X=Right). In other words, the adjectives with typically prenominal placement tend to appear more in the noun phrases situated to the left of their parent X, while the adjectives with typically postnominal placement tend to appear more in the noun phrases situated to the right of their parent X. Note that this bias does not seem to be an artifact of the corpus sample since, rather surprisingly, it arises in all five languages in our sample. There could be many potential explanations for this bias that are not directly related to DLM principle (e.g., the information status of the noun phrases with respect to their position in the sentence). As we discuss in the following, these data are nevertheless also consistent with the minimisation of X–N dependency, although not at the level of language variation, but at the level of grammar.

Discussion

Our analysis of the factors `Alpha` and `Gamma` has concluded that there is no clear minimisation effect for the adjective-noun dependency when $\alpha > 0$. Overall, it seems that the heavy adjective shift generalisation should be preferred over the adjective-noun dependency minimisation. There is, however, additional evidence for the adjective-noun dependency minimisation which comes from the fact that the order N YP Adj is very infrequent compared to the N Adj YP order. The heavy adjective shift principle cannot explain this asymmetry because both orders are postnominal. The

[15] The corresponding β values for the five languages, in their order in Table 4.8: 0.58^{***}, 0.34^{**} , 0.39^{***}, 0.44^{***}, 0.63^{***}

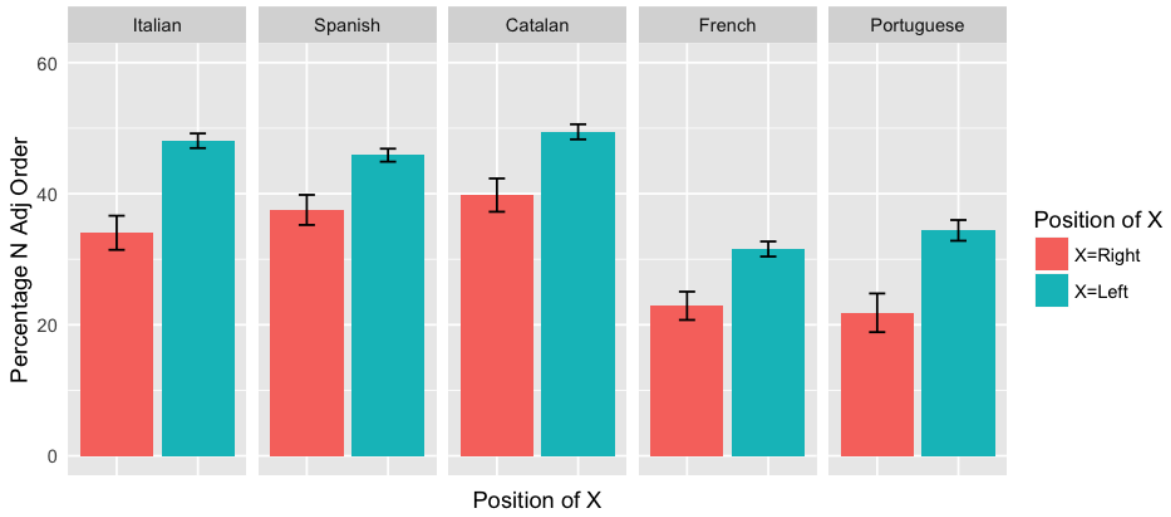


Figure 4.8: The percentage of postnominal adjectives for two positions of X: right (red bars) and left (green bars). Only noun phrases with simple adjectives and without other dependents Y are included.

dependency length minimisation of the noun-adjective dependency would clearly favor the order N Adj YP where the adjective phrase and the noun are adjacent. The phrase YP is typically longer than the phrase AdjP, therefore the order N AdjP YP will be preferred over the order N YP AdjP if we assume a simple DLM treatment parallel to the case of postverbal dependents V {XP, YP}.

In light of the evidence for the minimisation of the X–N dependency, we can address the postnominal preference with $\alpha > 0$, problematic for the DLM account, along the following lines. As we noted before, the position of X is predominantly on the left of the noun (80%). In this position, the minimisation of X–N implies the postnominal order favored by both `Alpha` and `Gamma`. We can speculate that this postnominal preference, occuring in the majority of the noun phrases, has been spread to the minority context where X appears on the right of the noun. In other words, we can assume that the DLM effect applies not at the level of variation — for each individual noun phrase and its dependency lengths — but at the level of grammar, similarly to the constructions analysed in Temperley (2007). Hawkins (1994, 2004) has claimed extensively that the DLM effects observed in the corpus data for inter-language variation are parallel to the tendencies observed at the level of grammars in

a typological sample. Similarly, the bias in the distribution of adjective types with respect to the position of X can be seen as a manifestation of a DLM principle at a level of abstraction higher than individual noun phrases.

At the moment, we do not distinguish between the cases of dependency length minimisation at the level of variation and at the level of grammar when we analyse corpus data automatically. The interaction between DLM effects and the syntax in general is an important topic for future research.

4.2.4 Summary

We can summarise the results of our statistical analyses as follows. The presence of a right nominal dependent Y is a highly significant effect, favoring consistently the prenominal placement of the adjectival modifier compared to its default position when Y is not present. Heavy adjective phrases containing pre-adjectival and post-adjectival dependents both favor postnominal placement compared to simple adjectives. The position of the parent of the noun phrase X has a global effect on the distribution of the adjectives favoring more prenominal adjectives when X is on the right of the noun and more postnominal adjectives in the opposite position.

Overall, we observe a complex case of variation with three dependencies that are minimised in the interaction with each other. These interactions are hard to analyse statistically in one model since the predictor variables are highly correlated. A simple global DLM model computing the sum of all dependencies in the sentence cannot capture these interactions.

4.3 Interaction of DLM and lexico-semantic factors in adjective variation in Italian

In this section, we take a closer look at the minimisation of N–Y dependency which, as we have shown above, has a consistent prenominal effect across all five Romance languages under investigation.

First, this preference has not attracted a lot of attention in the previous literature. In empirical work on French, Forsgren (1978) has noted the tendency for more balanced noun phrases which includes the preference for a more prenominal adjectival position when some other dependents are present in the noun phrase (confirmed later on a larger corpus by Thuilier (2012)). In the theoretical literature, however, such preference has not been mentioned and the N Adj YP order is considered, contrary to the observational facts, to be the default one (Laenzlinger, 2005).

Secondly, the minimisation of N–Y dependency occurs even when the adjective phrase consists of only one word (Figure 4.6). This is a surprising result given that the dependency which is minimised is very short. Since a processing explanation based on the memory storage constraints is improbable we would like to study the N–Y minimisation in more detail to understand the nature of the observed DLM effect. We focus on the adjective alternation in Italian and collect the statistics for adjective distribution subcategorised by the type of the right dependent YP. In fact, the YP notation we used in our analysis subsumes many types of syntactic phrases that can have different lexical and syntactic relation with the head noun.

We rely on the dependency annotation to extract additional information about the noun phrase such as the first word of the YP phrase and the head word of the YP phrase. We can use the first word of the YP phrase to subcategorise it into the phrase types. In our analysis, we retain only the most frequent types of the phrases: prepositional phrases (starting with a preposition) and relative clauses. We combine relative clauses which start with a relative pronoun (*che*) and reduced relative clauses without it. We identify the latter clauses as the dependents YP whose head is a verb (part-of-speech tag VERB) and which do not start with a preposition. Italian prepositions merge with definite articles into complex determiners such as *del* = *di* + *il*, *al* = *a* + *il*. We categorise these and other contracted variants of a preposition (e.g., *d'*, *ad*) under the same PP category labelled using the default form of the preposition.

Table 4.10 presents the percentages of prenominal, postnominal and post-YP placement of adjectives broken down for the most frequent types of PP phrases and relative clauses. For comparison, we also give the percentage of postnominal and prenominal placement when there are no right dependents in the noun phrase (first line of the

Type of YP		#	Adj N %	N Adj %	N YP Adj %
no YP		2767	46.7	53.3	—
PP					
	di	1161	66.8	28.9	4.3
	a	103	69.9	28.2	1.9
	in	96	63.5	32.3	4.2
	other	180	52.8	44.4	2.8
RelC		264	64.4	35.2	0.4

Table 4.10: The percentages of Adj N, N Adj and N YP Adj order broken down for the most frequent types of YP phrases in Italian.

table). We can observe that there is a substantial bias towards more prenominal orders for all types of PP phrases and relative clauses. We confirm this observation statistically by fitting a familiar mixed effect model with the type of YP (including the absence of YP) as fixed effect and the adjective lemmas as random effects (Table 4.11). The fitted coefficients show that all the YP phrases induce a prenominal preference but to a different extent. In particular, the prepositional phrases with *di*, the most common preposition in Italian, show the strongest tendency for preposing the adjective ($\beta = -1.49$). Relative clauses, on the other hand, show the weakest prenominal preference ($\beta = -0.61$). Overall the prepositional phrases seem to have a stronger preference for adjacency with the noun than relative clauses. Our result mirrors traditional grammatical analyses of the noun phrase where a noun and a prepositional phrase are placed closer in the syntactic structure than a noun and its relative clause (e.g., nouns can have selectional preferences for the preposition). The distributional data of adjective placement gives a new perspective on this relation.

The fact that prenominal placement is strongly favored by *di*-phrases warrants some attention. There are many noun phrases of N-di-N type which are analysed as lexical compounds and which we treat in the same way as other types of phrases in our analysis above. These include the expressions such as *casa di riposo* (elderly home), *colpo di fulmine* (love at the first sight), *punto di vista* (point of view) and others. This

YP type	Intercept	PP a	PP di	PP in	PP other	RelC
β	0.77^{***}	-1.42^{***}	-1.49^{***}	-1.01^{***}	-0.77^{***}	-0.61^{***}
	(0.10)	(0.28)	(0.11)	(0.29)	(0.20)	(0.18)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4.11: The statistical analysis of the effect of different YP phrases: $Order \sim YP\ type + (1 \mid Adj)$.

suggests that the bias observed can be due to a presence of some lexical constraints in addition to (or in substitution of) a more general DLM effect. To verify this lexical hypothesis we futher analyse the distribution of adjectives in *di*-phrases breaking them down into three categories: N-di-N phrases with a bare common noun following the preposition (*libro di vetta*, mountain-top book), N-di-N phrases with a proper noun (*libro di Maria*, book of Maria) and N-di-DP phrases where the preposition is followed by an article signaling a noun phrase (*libro della nonna*, book of the grandmother). The first type of noun phrases can be considered as fixed lexical units and an adjective is generally not allowed to intervene: **libro nuovo di vetta*.[16] In the other types of noun phrases this position is acceptable: *libro nuovo di Maria*, *libro nuovo della nonna*.

Table 4.12 gives the percentages of prenominal and postnominal adjectives when a right dependent of the noun is of one of the three types of *di*-phrases. As we hypothesised, there is a great number of prenominal adjectives in N-di-N phrases (74.7%) which must be due to their lexical compound status. Interestingly, the preference induced by other types of *di*-phrases remains strongly prenominal (it is statistically significant in a mixed-effect model, similar to the one presented in Table 4.11).

We can conclude that there is a strong lexical component in the minimisation effects that we have observed for N–Y dependency. It is, however, not purely lexical as manifested by the results for non-compound *di*-phrases and for the relative clauses. Moreover, if we assume that the adjective cannot intervene between the noun and its

[16]Note that such seemingly categorical constraints can be also sometimes violated, e.g.: *quel punto particolare*$_{Adj}$ *di vista*.

Type of di-N phrase	Example	#	Adj N	N Adj	N YP Adj
No YP		2767	46.7	53.3	—
Bare noun	*libro di vetta*	364	74.7	17.6	7.7
Det + N	*libro della nonna*	655	62.4	34.7	2.9
Proper noun	*libro di Maria*	107	67.3	31.8	0.9

Table 4.12: The percentages of adjective placement when YP dependent is a prepositional phrase with preposition *di* introducing a bare noun, a noun phrase (with a determiner) or a proper noun.

dependent Y for lexical reasons, that is, if N + YP form a lexical unit, then the order N YP Adj should be much more frequent than it is in the observed data (Tables 4.10, 4.12). There is an increase in N YP Adj order for N-di-N phrases (7.7%), however, it is a very small number compared to the default order of simple adjective plus noun phrases (53.3%). Interactions between lexical and syntactic properties behind the dependency length minimisation effects is an exciting topic for future research.

4.4 Conclusions and future directions

In this chapter, we have investigated the adjective-noun order variation in sentential context in Romance languages. To our knowledge, this is the first theoretical and empirical analysis of this construction in connection to the dependency length minimisation principles. In addition, as far as we know, this is the first cross-linguistic large-scale quantitative study of adjective variation across several Romance languages. Overall, we found that there is a significant influence of syntactic factors such as the lengths of dependencies on the adjective placement in Romance languages. The effects are very consistent across all five languages that we have studied: Italian, Spanish, Catalan, French and Portuguese.

Interestingly, we found that even very short dependencies such as the dependency between the noun and its dependent Y and the dependency between the noun and its head X are minimised to some extent. This is surprising given that DLM effects

are typically attributed to memory-related constraints on language processing. Note, however, that the effects we found cannot be purely lexical as we argued in Section 4.3. Alternatively, such short-range effects might be attributed to categorical, grammatical constraints, such as the fact that a complement should appear closer to the verb than an adjunct or that the adjective should appear closer to the noun than a relative clause (e.g., the order N RelC Adj is not possible). Our view is that there is a continuum of DLM effects which arise from a combination and interaction of various lexical, syntactical and processing constraints.

The approach we pursued here — adopting the most general DLM principle and relying on the dependency treebanks for syntactic analysis of sentences — proved to be a good way to approach a complex case of variation with many potential heterogeneous DLM effects. In future work, the application of this approach to the constructions that were not studied before will likely reveal related and new types of DLM effects. The first candidate constructions for an extended investigation could be the adjective and participle variation in Slavic languages, where, interestingly, an adjective or a participle with a complement can be found prenominally (seemingly violating the heavy adjective principle). Beyond the adjective variation, the construction of a related type is the alternation in the order of adverb and verb (Jackendoff, 1972; Alexiadou, 1997; Abeillé and Godard, 2003). Adverbs, e.g., in English or French, can appear in both pre-head and post-head position, similarly to adjectives in Romance, and the dependents of the verb, such as its object or indirect object, show a varying degree of adjacency effects similarly to the noun and its dependents (complements, PPs, relative clauses). The analyses of such constructions should lead to develop a more informed and accurate general DLM principle which could unify the effects of lengths of dependencies with lexical relations and which could explain explicitly the interaction of various dependencies.

The question of lexically-conditioned DLM effects is tighly connected with the assumptions on the underlying syntactic representation. As we argued in this chapter, an important advantage in using a pre-defined syntactic annotation is the generalisation of DLM effects across languages and constructions. We assumed, in fact, that word order is a (sophisticated) mapping between the flat hierarchical structures such as dependency trees to the linear arrangements of words. DLM effects are part of

the linearisation mechanism, conditioned on the underlying representation, but are not in any way coded in the input structure. Alternatively, richer hierarchical representations could contain more information for DLM encoding than flat dependency trees. For example, the difference in adjacency between the noun and its modifiers can be explicitly given by the phrase structure [[N Adj] PP] RelC]. If we assumed this syntactic analysis we could readily explain the preference for N Adj PP order versus N PP Adj order. On the other hand, this structure does not explain the fact that Adj N PP order is more frequent compared to Adj N order (when there is no PP). The question that arises naturally is whether we can induce the word order preferences that can be sometimes coded by deep(er) hierarchical syntactic structures from flat representations such as dependency trees. To answer this question we need a complete model of linearisation. The next chapter presents the first steps towards developing such model.

Chapter 5

A computational model of sentence linearisation and word order variation

The previous chapters presented two independent methods to analyse word order variation and DLM effects at two different levels of linguistic abstraction. In Chapter 3, we quantified the degree of DLM and word order variation at the language level. In Chapter 4, we studied in detail one particular word order variation construction and analysed its distribution independently of all other constructions in the language. Both studies leveraged dependency treebank resources to do the analysis but the computational methods employed were different: in Chapter 3, we quantified word order properties of a treebank as a whole based on high-level word order statistics and, in Chapter 4, we used a mixed-effect model to analyse and predict adjective placement for each sentence in a treebank. These two types of word order distributions at the two levels of representation — language level and construction level — are intrinsically related but have been analysed only separately in most previous work. In this chapter, we argue that they can be modelled and analysed jointly as part of a single sentence linearisation process.

The fundamental linguistic connection between word order distributions observed at the language level and at the construction level is that they are both generated by the same production system. During a speech act, an utterance is produced according to

some online processing mechanism that uses the grammatical knowledge of a speaker. The ensemble of utterances produced by speakers necessarily reflects properties and biases of the production mechanism and the grammar. These properties should be observed in language-level word order patterns and in preferences between word order options of one construction. If a pressure to minimise dependencies is a processing constraint influencing word order placement during production, it can potentially explain DLM effects observed both at the level of language and at the level of individual constructions and give them a unifying account.

Previous quantitative syntactic research on word order variation abstracts away from the properties and constraints of the language production mechanism. A typical analysis involves comparing two grammatical word orders and identifying and quantifying the effect of various properties on the preferences between the two (Gries, 2003; Bresnan et al., 2007; Thuilier et al., 2012) — this is our approach in Chapter 4 to study adjective alternation. One of the implicit assumptions behind such analyses is that the two alternative orders are fully constructed and available for comparison as complete sentences or phrases. This assumption is unrealistic from the production perspective: we do not plan complete utterances in advance, rather, they are constructed online as we speak. If memory retrieval mechanisms affect the choice of word order, we need an online model of word order production to incorporate them faithfully. In comprehension, memory constraints have been incorporated explicitly into a number of computational models including incremental probabilistic parsers (Vasishth and Lewis, 2006; Demberg and Keller, 2009a; Levy et al., 2009; Wu et al., 2010; Van Schijndel et al., 2013). For instance, difficulties in sentence processing, measured in self-paced reading time or eye-tracking experiments, were shown to emerge in a syntactic parser with explicit memory constraints (Demberg and Keller, 2009b). To our knowledge, there are no comparable computational psycholinguistic models developed to study word order and DLM effects in language production.

Language production is typically seen as a hierarchical incremental process which maps a mentally constructed "message" onto lexical items, their functional roles, a syntactic structure and, finally, a phonological realisation, that is, an utterance (Levelt, 1989; Bock and Levelt, 1994). Modelling computationally a complete language production system characterised by complex interactions and many levels of processing

is a challenging task. There exist several implementations of the complete language production system using the connectionist architecture (Dell et al., 1999; Chang, 2002; Chang et al., 2006). Due to computational limitations, these models are designed for simplified hand-crafted languages with small lexicons and very limited syntactic grammars including, e.g., only sentences with one clause. As such, these models cannot be used for large-scale analysis of naturally occurring word order variation data. Moreover, they are designed for one language (English) and cannot be applied cross-linguistically.

In this work, we do not attempt to model the mechanism of language production in its entirety. Instead, we focus only on the level of processing relevant to word order. We assume a serial architecture where, at the moment of uttering a sentence, the underlying syntactic representation has already been generated by a speaker, and we are interested in the process of producing a word order given the underlying syntactic structure, a process known as *sentence linearisation*.[1] In the NLP literature, models of sentence linearisation have been previously developed as part of natural language generation (NLG) systems. Since these systems are devised for practical NLP applications, they are robust and language independent but, at the same time, they do not bear directly on the study of human language production. In this chapter, we propose the first statistical system which is explicitly designed to be a plausible psycholinguistic model of sentence linearisation. Developing this system is a first step towards studying word order and DLM effects at several linguistic levels as part of one language production process.

This chapter is organised as follows. We start by reviewing the relevant language production literature and motivating the cognitive requirements for a sentence linearisation system. These include online, word-by-word processing and a generative probabilistic architecture (Section 5.1). Next, we present the architecture of our system which generates one word at a time based on a score function which combines the probability of the subtree uttered so far and a future score for generating the remaining words (Section 5.2). To assess the limits of the incremental greedy search, we evaluate our system on four languages and show that it can reach good performance,

[1] Sentence linearisation can be seen as a reverse of sentence processing, where the input is the words and their order and the output is the syntactic structure.

compared to a state-of-the-art statistical linearisation system, by predicting only several words at each step (Section 5.3). In the last part of this chapter, we show how our basic linearisation system can be extended to model explicitly the choice between two alternating orders using a re-ranking function (Section 5.4). This mechanism also allows us to introduce an efficient way to condition online word order choices on dependency length factors. In this way, we integrate word order variation preferences and DLM pressures into a more general word order production system.

5.1 Background and motivation

This section motivates several cognitive constraints on the architecture of a sentence linearisation system imposed by experimental evidence from language production. We also give an overview of previously proposed statistical models of language production and sentence linearisation.

5.1.1 Language production and cognitive basis for a sentence linearisation system

A dominant view in the language production literature is that the process of production consists of a series of stages corresponding to specific levels of linguistic representation (Fromkin, 1971; Garrett, 1988; Bock and Levelt, 1994). These serial models assume two main separate production processes: grammatical encoding and phonological encoding. We are interested here in the grammatical encoding which comprises the selection of lexical items and construction of the syntactic structure. The input to grammatical encoding module is assumed to be some form of a mental message — a representation of meaning formed by a speaker. For example, when a person is given a picture, she observes it and creates a mental message consisting of the concepts depicted in the picture. She does not necessarily choose at this point concrete lexical items for these concepts. The grammatical encoding is separated further into two modules: functional and positional (Garrett, 1988; Bock and Levelt, 1994). Functional processing involves lexical selection (e.g., the choice of lemmas

for the concepts) and the assignment of grammatical roles for the elements of the main event frame (e.g., a subject or an object). Positional processing consists in linearising the lexical items given the functional structure and enriching them with correct grammatical inflection. We will also adopt a strong assumption that the processing levels of language production are executed in a pipeline (Bock and Levelt, 1994; Vigliocco and Nicol, 1998). In other words, processing at each level is influenced only by the information provided by the level directly above it.

The division between functional and positional levels of processing implies that word order is selected after the assignment of grammatical structure. In Chapter 2, we have motivated a similar view on separating hierarchical syntactic representation and linear precedence constraints from a theoretical syntactic perspective. The evidence for this distinction in production comes from the exchange errors which were shown to happen with structurally-related but not necessarily linearly adjacent phrases (Garrett, 1980). Moreover, priming experiments showed that grammatical functions and linear orders induce separate priming effects (Bock et al., 1992).

In this work, we are interested in modelling the linearisation part of the positional processing module of language production. Given the assumptions outlined above, this corresponds to modelling the mapping of the underlying syntactic tree structure onto its linearisation. Following NLG notation, we will refer to this task in the language production pipeline as sentence linearisation.

Natural language production and its processing modules exhibit some properties which should be reflected in our model of sentence linearisation. Production is characterised, in particular, by simultaneous planning and execution of utterances (Ferreira and Swets, 2002; MacDonald, 2013). Speakers do not plan their complete utterances in advance: if this were the case, then we would expect long pauses between utterances (Ferreira, 1991); instead, it is well documented that disfluencies, lengthening of words and interjections occur mid-utterance in an attempt to gain extra planning time (Fox Tree and Clark, 1997). However, there is also clear evidence that speakers plan more than one word at a time and that sometimes language production requires the retrieval of non-local syntactic context, e.g., the retrieval of the postverbal object in English before the utterance is started (Ferreira and Swets, 2002; Meyer,

1996). In the language production literature, this general property of the production system to interleave planning and execution is referred to as *incrementality*.

The incrementality of serial language production means that generation at all levels of representation (functional, positional, phonological) proceeds online and in parallel. The formulation of the sentence linearisation task which is adopted in NLG work simplifies this picture: the input syntactic representation driving the choice of word order is assumed to be available in its entirety before the linearisation starts. The model we propose does not make this strong input assumption: the linearisation proceeds recursively by subtrees, or word order domains (Reape, 1994), formed by the head and its immediate children; we assume that these are the only elements accessible during each word ordering decision. For example, to start a sentence, only the head verb and its immediate dependents such as the subject head word and the object head word must be available. This corresponds to the assumption that the overall predicate structure of the sentence has been decided before speaking begins, which is a much weaker assumption than the assumption that the whole syntactic structure is fixed beforehand. The syntactically-local linearisation process outlined above is further supported by the evidence from Bock and Cutting (1992) who found that noun phrases in the embedded clauses do not induce agreement errors in the main clause. This suggests that the main clause and the embedded clauses are produced, to some extent, independently.

In the context of computational modeling of sentence linearisation, we will also use the term incrementality to refer to the ability of the system to generate a sentence as an online process, left-to-right and word-by-word. This notion is similar to the notion of incrementality in sentence processing. Analogously to processing, the decision about generating the next word is conditioned on the previously generated words. An important difference between linearisation and parsing mechanisms is that the former has access to the underlying syntactic structure. The basic version of our linearisation model is a purely incremental model that generates one word at a time. It is the least cognitively demanding model since only the choice of one word is entertained at each point in time, and a small local space of possible continuations is explored. We also compare this minimal model to a more powerful system which produces several words at a time. At each step, it explores a larger space of possible

continuations and therefore requires more processing resources than the simpler word-by-word system.

Finally, speakers are known to have probabilistic syntactic and lexical knowledge (Manning, 2003; Bresnan et al., 2007) and they employ this knowledge during syntactic processing and production (Stallings et al., 1998; Hale, 2001; Jurafsky, 2003). Moreover, production and comprehension (including linearisation and syntactic processing) are tightly interleaved: during comprehension, listeners make predictions about upcoming structures which are claimed to originate from a simultaneous production process mirroring that of the interlocutor (Pickering and Garrod, 2007, 2013). The probabilistic predictions speakers make should, therefore, be a part of the production system. To model this aspect of production, we condition the choice of the next generated word in our system on the probabilities of a generative dependency grammar learned from a treebank.

5.1.2 Computational models of language production and sentence linearisation

The most developed model of the complete language production process — mapping a message to the utterance — is the connectionist dual-path model (Chang, 2002; Chang et al., 2006; Chang, 2009). It is characterised by a recurrent neural network predicting the next word given the previous word (sequencing system) and an additional "path" encoding the input message and the relation between concepts, functional roles and lexemes. The syntax is not available as input for this model, only a set of concepts defining the semantic message of the sentence. Instead, syntactic information is assumed to be learned implicitly by the recurrent neural network. One of the main difficulty for these models lies in encoding semantic and lexical information. For example, the neural network encodes each of the possible lexical units (verbs, nouns, adjectives) as a distinct neuron of the lexical layer. This means that the lexicon should be pre-coded before the training of the model which is unrealistic for a natural language. Consequently, connectionist production models are typically trained on a simple lexicon of only several hundreds of units. By a way of simulation, the

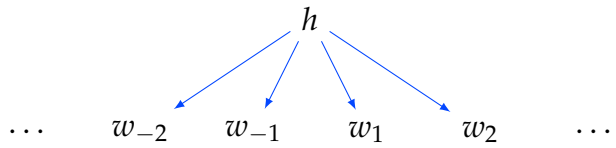


Figure 5.1: A subtree headed by h which is linearised by the generative process. The subtrees headed by each w_i are linearised recursively.

dual-path model was applied on selected cases of word order alternation, including the modelling of heavy-NP shift in English and Japanese (Chang, 2009). The results suggest that heavy-NP shift can originate implicitly through the general learning biases of the computational model as opposed to independent DLM effects. However, the limited applicability of the dual-path model makes it unsuitable for studying word order variability and DLM effects at large scale and across different types of constructions and languages.

The sentence linearisation part of the language production process has been modelled more extensively in the statistical NLG literature (Filippova and Strube, 2009; Bohnet et al., 2010; Wang and Zhang, 2012; Bohnet et al., 2012; Liu et al., 2015; Puduppully et al., 2016). The shared task on sentence linearisation (Belz et al., 2011), referred to as surface realisation, aimed to provide a standardised reliable comparison of systems and proposed an evaluation on the same input data based on dependency annotation. Among the sentence linearisation models that have been proposed, the probabilistic generative system of Futrell and Gibson (2015) and the transition-based ZGen (Liu et al., 2015; Puduppully et al., 2016) are the ones most directly related to our work.

Futrell and Gibson (2015) applied top-down generative dependency parsing models (first introduced by Eisner (1996) and subsequent work) to the sentence linearisation task. The model of Eisner (Model C in his paper), is a simple generative model which defines how immediate dependents are generated conditioned on the head h (Figure 5.1). The children on the left and on the right of the head are generated independently, based on a head-outward Markov process. In other words, the i-th child w_i (w_{-i}) is generated conditioned on the previously generated child w_{i-1} (w_{-i+1}) with the probability $p_R(w_i \mid h, w_{i-1})$ for the children on the right of the head and the probability $p_L(w_{-i} \mid h, w_{-i+1})$ for the children on the left of the head. The

probability of the subtree formed by the head and its immediate children linearised as $w_{-l} \dots w_{-1} h w_1 \dots w_r$ equals therefore $\prod_{i=1}^{r} p_R(w_i \mid h, w_{i-1}) \cdot \prod_{i=1}^{l} p_L(w_{-i} \mid h, w_{-i+1})$. The overall probability of the tree is obtained recursively by multiplying probabilities for each subtree, starting from the root node. Given this generative model one can compute the probability of a particular order $w_{-l} \dots w_{-1} h w_1 \dots w_r$ given the unordered set of children w_i under the head h. Futrell and Gibson (2015) propose a dynamic algorithm to compute these values efficiently for all possible orderings. They evaluate various versions of the generative dependency model with different smoothing methods for probability estimation and on a number of languages. They also test an extended Eisner model which includes larger conditioning context of n-grams of size 3, i.e., $p_R(w_i \mid h, w_{i-1} w_{i-2})$. The authors report human evaluation results: acceptabilty of up to 3.6/5 (original sentences have average acceptability of 4.5/5) and proportion of reordered sentence judged to have the same meaning as the original English sentence up to 85%. Unfortunately, they do not conduct any direct comparisons of their models with the previously proposed linearisation systems and these numbers are hard to place in context. The BLEU score on their English treebank reported is only 57.7 which is substantially below the best results in the Surface Realisation shared task Belz et al. (2011), reaching 89 BLEU, though these numbers are not directly comparable.

Importantly, the generative model used in Futrell and Gibson (2015), as well as other previously proposed top-down statistical linearisation models (Guo et al., 2011), requires a global search to find the most probable word order. This implies a production process where planning of the complete word order happens before the production of the first word. In comparison, we propose to model language production incrementally as an online procedure generating the next word using local decisions.

The state-of-the-art system of Liu et al. (2015); Puduppully et al. (2016) is another adaptation of a parsing system to the task of sentence linearisation and one of the few systems freely distributed online.[2] It is based on the transition-based parsing architecture (similar to MaltParser, Section 3.2.1) as opposed to the generative parsing architecture and is incremental in the sense that it chooses the next word based on

[2] `https://github.com/SUTDNLP/ZGen`

previously chosen words. To our knowledge, it is the only incremental architecture proposed so far for the sentence linearisation task. Despite this, ZGen cannot be taken as a model of the cognitive process of linearisation. First, it uses a beam search with the default number of 64 hypotheses. It means that a list of partial hypotheses of the word order (fixing the words from 1 to k) is kept in memory and the next decision consists in expanding these partial sequences of words (fixing the words from 1 to $k+1$) and constructing a new list of best-scored hypotheses. In addition, ZGen uses a carefully designed set of discriminative lookahead features which include combinations of the word to generate next with its sister and child nodes not ordered yet. Both these mechanisms allows the model to explore a large portion of the search space in a non-incremental way. Secondly, as a transition-based system, ZGen does not have a probabilistic interpretation. To predict the next word (or, more precisely, the next transition), the algorithm uses a discriminative classifier on the state features which do not explicitly include syntactic and lexical frequencies. Overall, these properties limit the applicability of ZGen for cognitive modeling. On the other hand, the properties such as a large beam make it a strong upper bound for performance comparison with a simpler and incremental model.

5.2 Architecture of the sentence linearisation model

The input of the sentence linearisation model is a hierarchical structure of an utterance which in this work is taken to be an unordered dependency tree. This choice allows us to use familiar dependency treebank resources to evaluate our sentence linearisation system and to compare its performance with other sentence linearisation systems developed for dependency trees in the NLG literature. As we discussed in Chapter 2, the characteristic properties of the UD dependency respresentation such as cross-linguistically universal dependencies relations and the priority given to the relations between content words makes unordered UD trees well suited as the input for the mapping between syntax and word order.

Specifically, the input for our sentence linearisation system is an unordered dependency tree consisting of the words (the tree nodes) and the grammatical relations

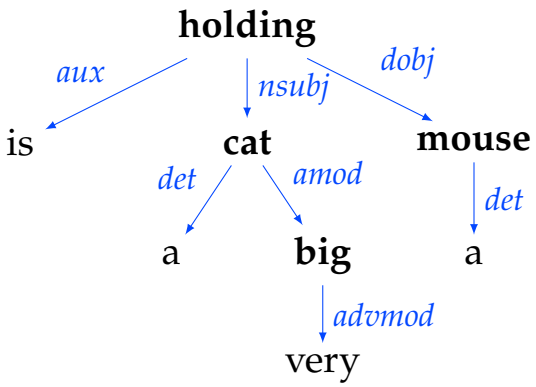


Figure 5.2: An unordered dependency tree representing the sentence *a very big cat is holding a mouse*.

between them (the dependencies). Figure 5.2 shows the unordered dependency tree of an example sentence *a very big cat is holding a mouse* which we will use to illustrate the sentence linearisation procedure.

In a sum, the linearisation procedure consists in traversing the tree in a top-down fashion and generating the order of the immediate children of each node (with respect to each other and with respect to the head). This basic recursive procedure is very similar to many generative models of tree structures, starting from the phrase structure elaboration process proposed for production by Yngve (1960). The distinguishing property of our model is that the next node to be ordered is chosen greedily based on the previously generated words and a score function which incorporates limited lookahead on the nodes not ordered yet. Below, we describe in detail the search procedure and the score function that we use.

5.2.1 Top-down recursive procedure

The diagram in Figure 5.3 illustrates the tree traversal and the linearisation for the sentence *a very big cat is holding a mouse*. The scheme presents an extended version of the dependency tree, where each head appears twice – as a head of a subtree (i.e., a non-terminal, in bold) and as a leaf (i.e., a terminal word). The bulk of the work lies in ordering correctly the nodes of the same subtree, i.e., the head and its immediate

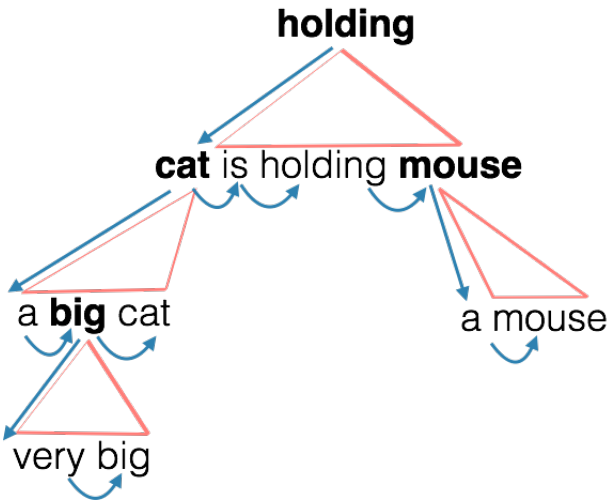


Figure 5.3: An example of the top-down left-to-right linearisation procedure for the dependency tree representing the sentence *a very big cat is holding a mouse*. The words in bold are non-terminal nodes in the tree.

children. For example, incremental ordering of the phrase *a very big cat* proceeds as follows, given an oracle score function predicting the correct next word. We start by choosing the first word among all the immediate children of *cat* (nodes *a* and *big*) plus the word *cat* itself. Assume the choice falls on *a*. Since *a* doesn't have children, we generate it and proceed to choose the next node among the remaining nodes (*big* and *cat*). Now we choose the node *big*, but we don't immediately generate its word. Instead, we expand the subtree headed by *big* (*very big*) recursively and generate all its nodes applying our left-to-right linearisation based on the score function predicting the next word. After spelling out all words in the subtree *very big*, we output the last word *cat*.

More formally, our greedy recursive linearisation procedure starts from the root node of the unordered dependency tree and proceeds as follows (Algorithm 1):

1. Form a set of nodes D_h by combining the head node h and its immediate children $\{n_i\}$ (line 2).

2. Incrementally order the nodes in this set from the leftmost node to the rightmost node starting from a special symbol n_0 (lines 3–13):

a) Choose the best next node n_k from the set D_h according to the score function of the prefix $p = n_0, n_1, \ldots, n_{k-1}, n_k$ where $n_0, \ldots, n_{k-1}$ are previously generated nodes from D_h (line 5).

b) If the node n_k does not have any children or is the head h, produce the chosen word (line 7); otherwise linearise the subtree headed by this node recursively (line 9).

c) Continue until all the nodes are output.

Algorithm 1 Incremental linearisation procedure

```
 1: procedure LINEARISE(h, t)
 2:     D_h ← {h, CHILDREN(h, t)}
 3:     p ← n_0
 4:     while D_h not empty do
 5:         next ← arg max SCORE(D_h, p, next)
 6:         if next is h or CHILDREN(next, t) is empty then
 7:             yield next                          ▷ produce the word next
 8:         else
 9:             LINEARISE(next, t)
10:         end if
11:         p ← APPEND(p, next)
12:         D_h ← D_h \ next
13:     end while
14: end procedure
15: LINEARISE(ROOT, t)                  ▷ t is the input unordered dependency tree
```

Our procedure orders subtrees independently from each other which implies, among other things, that nodes in a subtree form a contiguous string. In other words, the linearised dependency trees produced by our linearisation model will always be projective. It is an important and interesting question how to model non-projective linearisations but it is out of scope of the current work.[3]

The basic procedure outlined above assumes that the choice of the next word is greedy. From the processing perspective, this means that only one partial hypothesis should be maintained during production and the words could be, in principle,

[3]Section 6.2 entertains some ideas for future work in this direction.

uttered straight after being chosen by the linearisation module. A simple greedy model of this type would also require a small amount of memory and computational resources during processing. We can conceive a more powerful model in two ways: first, more than one partial linearisation hypothesis could be maintained by the speaker or, secondly, the placement of more than one word could be chosen at each linearisation step. In the sentence comprehension literature, both serial and parallel processing mechanisms have been considered cognitively plausible (Boston et al., 2011). Production is, however, different from comprehension in this respect. If several hypotheses are constructed during the linearisation process, the speaker cannot start phonological production if these hypotheses differ in the order of the very first words or phrases. In other words, maintaining several linearisation hypotheses is equivalent to optimising word order globally (in the worst case).

Here we explore a second variant of the linearisation model which orders several words at a time. This model requires more computational time to produce the next word since a larger number of possible continuations is evaluated compared to predicting only one word (e.g., $m \cdot (m-1) \cdot \ldots \cdot (m-l)$, where m is the number of words to order and l is the number of words to predict next). Crucially, however, linearisation proceeds in the same serial greedy manner commiting to one partial hypothesis from the beginning.

The recursive linearisation procedure is straightforward. The main challenge is to design an efficient and accurate function which determines which node is generated next. We propose such probabilistic scoring function in the next section.

5.2.2 Probabilistic score function

Our starting point for ordering the words in every set D_h is a probabilistic generative model akin to an n-gram language model that estimates the probability of the prefix $p = n_0, n_1, \ldots, n_k$ (where $n_i \in D_h$) as a product of the conditional probabilities of each node $P(n_i \mid h, n_0 \ldots n_{i-1}), i = 1, \ldots k$. We further factorise these probabilities into direction and n-gram probabilities similarly to the generative model proposed

by Eisner (1996) for dependency parsing and adopted by Futrell and Gibson (2015) for sentence linearisation.

The main difference compared to the previously proposed generative dependency models is that we do not directly use the probability $P(p)$ as our score function. Instead, we define the score of a prefix p as the product of its generative probability, factorised as the product of the probabilities $P(n_i)$ for each of the nodes $n_i \in p$, and the *future score* — the score of the nodes $n_j \in D_h, n_j \notin p$ that have not been ordered yet:[4]

$$Score(p) = \prod_{n_i \in p} score(n_i, p) \cdot \prod_{n_j \notin p} score_f(n_j, p) \tag{5.1}$$

The first part of the $Score(p)$ is defined as follows:

- if $n_i = h$, $score(n_i, p) = P(n_i \mid n_0 \dots n_{i-1})$
- if $n_i \neq h$:
  - if $h \notin \{n_1 \dots n_{i-1}\}$, $score(n_i, p) = P(n_i \mid h, n_0 \dots n_{i-1}) \cdot P(left \mid n_i, h)$
  - if $h \in \{n_1 \dots n_{i-1}\}$, $score(n_i, p) = P(n_i \mid h, n_0 \dots n_{i-1}) \cdot P(right \mid n_i, h)$

The probability of generating a child is factorised as the probability of generating it on the left or on the right of the head (direction probability) and the probability of generating it given the previously generated children (n-gram probability) and the head.

The second component of $Score(p)$ is a score which estimates an upper bound on the probability of a linearisation of the rest of nodes in D_h, if we commit to generating p. It is defined as:

- if $n_j = h$, $score_f(n_j, p) = 1$

[4] The use of future score in linearisation is, perhaps, reminiscent of the similar use of future scores for efficient A* search in phrase-based statistical machine translation. The conceptual relation between language production and machine translation processes is evident. In early NLP work, language generation was considered as the last step in the pipeline of machine translation (Yngve, 1960). Early statistical NLG work (Langkilde and Knight, 1998; Bangalore and Rambow, 2000) uses n-gram scoring over word lattices produced by generation grammars, similarly to some decoding techniques used in MT.

- if $n_j \neq h$:
  - if $h \notin p$, $score_f(n_j, p) = \max(P(left \mid n_j, h), P(right \mid n_j, h))$
  - if $h \in p$, $score_f(n_j, p) = P(right \mid n_j, h)$

It can be seen as an upper bound since it is equal to the probability $score(n_j, p)$ where n-gram probability $P(n_j \mid h, n_0 \dots n_{j-1})$ is equal to 1. The future score is an important mechanism to constrain the space of hypotheses for the generative model. It is designed to take into account the fact that the generation of the head node defines the relative placement of all the children, including the ones which have not yet been generated. If the prefix already contains the head word then we know that all the following nodes will be on the right of the head and the future score is equal to $P(right \mid n_j, h)$. If we do not know yet where the head is, then node n_j can end up on its left or on its right and we use the maximum of the two probabilities as the upper bound on its potential score.

The future score providing limited lookahead for the model is necessary since we want to obtain an accurate greedy linearisation procedure. For generative models without lookahead, a large beam is typically required to find good final hypotheses (Henderson, 2003; Titov and Henderson, 2010). The future score is also a plausible addition from the perspective of the input representation. Indeed, we assume that all children of the head in the set D_h are known to the speaker.

Given a previously constructed partial order p, a next node (Algorithm 1, line 5) is chosen according to the best-scored new prefix with an appended node $\langle p, n_k \rangle$:

$$next = \arg \max_{n_k \in D_h \setminus p} Score(\langle p, n_k \rangle). \tag{5.2}$$

In case of models predicting m words at a time, this step is modified to evaluate all possible extensions of p of length m:

$$next = \arg \max_{<n_1, \dots n_m> \in S_m[D_h \setminus p]} Score(\langle p, n_1, \dots n_m \rangle). \tag{5.3}$$

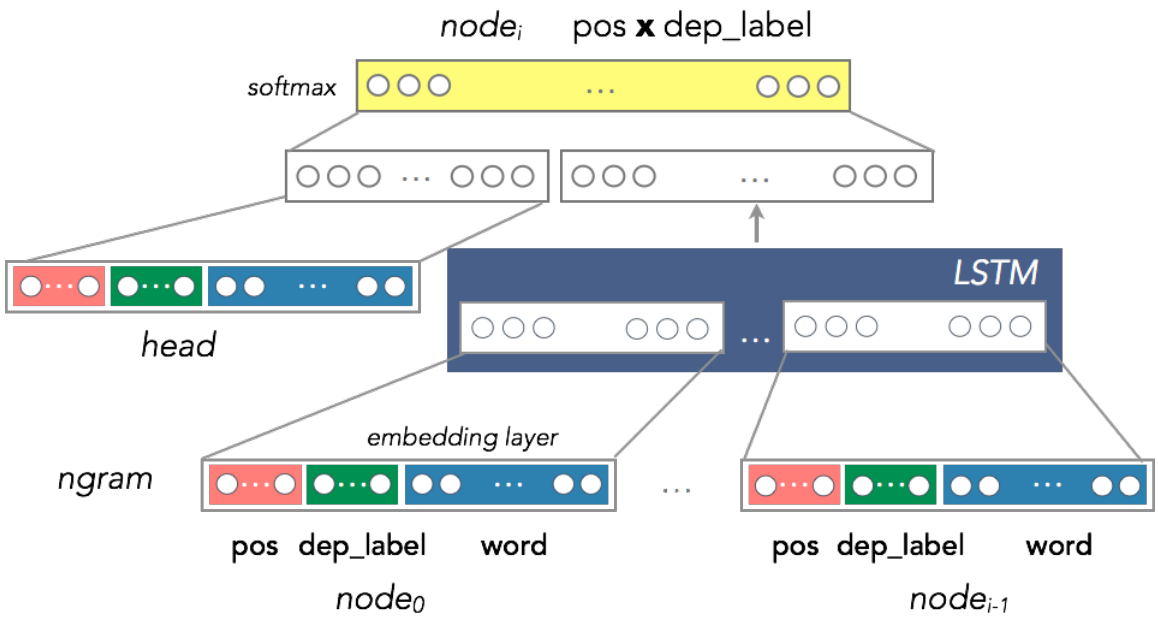


Figure 5.4: Neural network architecture for estimation of n-gram probabilities.

For example, a model predicting three words at a time evaluates all permutations $\langle n_1, n_2, n_3 \rangle$ for all possible $n_1, n_2, n_3 \in D_h \setminus p$ and chooses the one which produces the highest score.

5.2.3 Estimation of probabilities

Arc-direction probabilities We compute arc-direction probabilities using frequency counts for all head-child pairs defined by part-of-speech tags and dependency labels. Since the space of observations is relatively small we do not apply any smoothing to this set of parameters.[5]

Estimation of n-gram probabilities We estimate the n-gram probabilities using a neural network to incorporate lexical features and to deal with sparseness. In the preliminary experiments, we also experimented with unlexicalised frequency-based

[5]Preliminary experiments did not show significant improvement when the arc-direction probabilities were estimated using a neural network or by adding additional lexical features.

estimation, similarly to the estimation of the arc-direction probabilities. We report only the results of the neural network approach since it consistently outperformed the frequency-based approach. Also, while the results of the latter approach change significantly based on the input features of the nodes (PoS tags or dependency labels or both), the neural network functions as a feature selection mechanism and allows us to avoid the additional manual feature engineering.[6]

Neural networks are now habitually used to produce accurate standalone and integrated language models and have been shown to perform better or on par with the classical language models built using maximum likelihood estimation and smoothing techniques (Bengio et al., 2003; Mikolov et al., 2010; Collobert et al., 2011). Recurrent neural networks and, in particular, Long Short-Term Memory (LSTM) networks have proved to be the best-suited architectures for the language modelling task (Sundermeyer et al., 2012; Jozefowicz et al., 2016). Feed-forward and recurrent neural networks are also now commonly used as building blocks of parsing systems (Henderson, 2003; Titov and Henderson, 2007; Chen and Manning, 2014; Weiss et al., 2015; Dyer et al., 2016). One of such building blocks, known as word embeddings, allows learning a representation of a category in a high-dimensional discrete space (for example, a vocabulary of word forms or PoS tags) as a low-dimensional continuous vector.

Our lexicalised model is schematically presented in Figure 5.4. We experiment with both lexicalised n-gram probabilities (input and output include the word feature) and unlexicalised n-gram probabilities (input and output include only PoS tag and dependency label features). The architecture of the two types of models is the same apart from the input and output layers.

The neural network encoding of the input to the output proceeds as follows. First, we map PoS tag, dependency label and word features of nodes $n_1, \ldots, n_{i-1}$ in the n-gram to dense embedding vectors, following Chen and Manning (2014) (among

[6]Neural networks are very powerful machine learning models which allow to fit complex functions from the input to the output based on labelled training data. However, the use of neural networks in this work is limited to estimating only simple n-gram probabilities and an extensive background in neural networks is not required for understanding of the following discussion. We address the reader to the textbook of Goldberg (2017) for a detailed introduction to neural networks for NLP.

many others) which use similar representations for parsing. We use vectors of size 32 for PoS tag and dependency label embeddings and vectors of size 64 for word embeddings. The combined vector of embeddings for each of the n-gram nodes is then passed to an LSTM layer with 128 hidden units. In addition, we incorporate the features of the head word in our network. The combined output of the LSTM layer and head embedding is passed to a fully-connected hidden layer with 128 non-linear (ReLU) units which is then mapped to the output softmax layer. The softmax layer is constructed to output the probability of observing a node n_i, defined by its output features: PoS, dependency label and word. The output space of the unlexicalised model is limited to all possible combinations of PoS tags and dependency labels. For lexicalised models, we use a small 500-word vocabulary to avoid efficiency and sparsity issues. The softmax output vector gives, therefore, probabilities of the 500 most frequent words in the training set. The words which are less frequent receive the probability based on their PoS tag and dependency label as in the unlexicalised model. Note that traditional n-gram language models use much larger vocabularies but require millions of words in a training corpus to achieve a robust estimation of the probabilities. Some sentence linearisation models use additional raw text data to train a lexicalised language model (so-called pre-trained word embeddings). In this work, we focus on the performance of a sentence linearisation system given only a treebank, without any additional resources.

We implement our neural network models using the Keras package with Theano backend (Chollet et al., 2015).[7] We use adaptive gradient descent optimisation known as Adam (Kingma and Ba, 2014) to fit the parameters of the model to the training data.

[7] `https://keras.io/`

5.3 Evaluation of the basic sentence linearisation model

In this section, we evaluate and analyse the linearisation algorithm presented above. Our first goal is to demonstrate the feasibility of the proposed incremental procedure. We want to understand whether the model can find globally-coherent orders by relying on greedy and, to a large degree, local decisions. We compare our model to the state-of-the-art sentence linearisation system, which provides a strong upper bound on non-global linearisation. Our point of comparison is the latest version of ZGen (Puduppully et al., 2016). As discussed in the Section 5.2, it is a powerful system exploring many combinations of rich non-local features to find the best linearisation. It achieved state-of-the-art results on the surface realisation shared task data in English improving on the results obtained during the task (Belz et al., 2011). We use ZGen in its default configuration which sets the size of the beam of its search algorithm to 64.

Our evaluation also aims to establish the general relation between the word order properties of a language and linearisation performance. We evaluate our system on four languages with relatively diverse word order properties: English, Italian, Persian, and Russian. English has relatively fixed word order, Italian and, especially, Russian show more word order flexibility and Persian is interesting because it is an SOV language, while the other languages have SVO order.

5.3.1 Data

The dependency treebanks of our four languages come from the Universal Dependencies (v1.3) annotation project (Nivre et al., 2016). We strip the sentences of their punctuation marks, since we are interested in language production in general and not necessarily in the production of written text only.[8] We use the word, coarse universal

[8]Contrary to the experiments in Chapter 3, we do not remove sentences which contain punctuation from our sample but simply strip them of the punctuation marks. For parsing, punctuation marks such as parentheses provide important cues about the phrase structure; a sentence with stripped

Language	Training set	Dev. set	Test set
English	180K	22K	22K
Italian	219K	9.5K	9.7K
Persian	110K	14.5K	14.5K
Russian	250K	25K	25K

Table 5.1: Sizes of the training, development and testing sections of the treebanks.

PoS tag, and dependency label information provided in the treebanks for the input dependency trees. Contrary to the Surface Realisation shared task (Belz et al., 2011), we do not pre-process multi-word expressions and proper names as single tokens.[9] Because of the different pre-processing and the use of a different corpus, our results cannot be directly compared to the previous results on English. We provide a fair comparison with previous work by reporting the performance of ZGen which was shown to out-perform the systems participating in the Surface Realisation task. To our knowledge, Italian, Persian and Russian languages have not been previously used for sentence linearisation evaluation. Our results on these languages therefore give first baselines for future work.

We train our linearisation model and ZGen on the training sets of the treebanks and analyse their performance on the development and test sets. For English, Italian and Persian we use training-development-test splits provided in the UD distribution. For Russian, which is a very large treebank, we used only 250'000 words for training data and 25'000 words for testing and development sets to keep the sizes of the datasets comparable across languages. The exact sizes of our datasets are reported in Table 5.1. To analyse in detail the performance of our system and the effect of its parameters we performed the majority of the experiments on the development sets.

punctuation marks can be therefore hard to parse correctly. In comparison, a linearisation algorithm has access to the underlying syntactic tree and should rely less on punctuation marks for producing the correct order. We assume therefore that the removal of punctuation marks does not affect the sentence linearisation task significantly.

[9]Despite the similar annotation guidelines in four different languages, such pre-processing would require language-specific manual verification which we could not perform.

5.3.2 Evaluation measures

We evaluate the order that is produced by the systems against the gold-standard original order using BLEU score — an n-gram precision measure commonly used in machine translation (Papineni et al., 2002). A BLEU score captures the surface similarity of the output order and the original order of words. As in machine translation, the conceptual problem of BLEU is that it cannot capture the semantic equivalence between the two outputs. For the task of sentence linearisation, it means that BLEU will find grammatical word order alternations of the gold order to be incorrect (n-gram precision will be less than 100%). Despite this issue, BLEU is standardly used for evaluation in the task of sentence linearisation and, more generally, for evaluation of natural language generation (Belz et al., 2011). First, BLEU scores were shown to correlate well with other automatic measures of generation accuracy adopted from machine translation such as NIST, TER and METEOR on English and German (Reiter and Belz, 2009; Cahill, 2009). Secondly, the ranking of the systems obtained through human judgements on clarity, readability and meaning similarity correspond well to the ranking given by automatic evaluation scores (Reiter and Belz, 2009; Belz et al., 2011), leading to the suggestion that: "it may be appropriate to use existing automatic metrics (with caution) to evaluate the linguistic quality of generated texts" (Reiter and Belz, 2009, p. 555). Based on these observations, we adopt BLEU as our primary measure of the performance of sentence linearisation systems. We leave a more detailed human evaluation analysis for future work.

In addition to BLEU, we report arc direction accuracy for the output order, i.e., the percentage of children that are placed correctly with respect to their heads. This measure is particularly interesting for us in relation to the treebank-level analysis of the arc direction entropy presented in Chapter 3. Note that BLEU scores do not distinguish between all cases where a head H and its dependents (D_1, D_2) are ordered incorrectly. Consider for example the gold order $H\ D_1\ D_2$ and two output orders $H\ D_2\ D_1$ and $D_1\ H\ D_2$. Both outputs will have a similar n-gram precision error between H, D_1 and D_2 phrase boundaries but the first order has correct head-child dependency directions. Conversely, two orders can have the same arc-direction entropy (e.g., $H\ D_1\ D_2$ and $H\ D_2\ D_1$) but different BLEU scores. The arc direction accuracy measure seems

	English		Italian		Persian		Russian	
Systems	BLEU	arc dir %	BLEU	arc dir %	BLEU	arc dir %	BLEU	arc dir %
ZGen	84.8	97.1	82.1	95.4	80.9	98.1	68.6	91.4
Unlex 1w	66.7	93.1	65.1	90.0	62.5	94.3	49.8	83.7
Unlex 3w	77.3	95.6	71.9	92.3	75.4	97.2	58.5	87.7
Unlex 5w	78.9	95.9	73.5	92.8	77.2	97.8	60.6	88.6
Lex 1w	66.9	93.4	66.4	90.7	65.2	94.6	51.5	84.6
Lex 3w	78.8	95.7	73.6	93.0	75.9	97.4	60.1	88.1
Lex 5w	80.5	96.0	75.4	93.4	78.4	97.9	61.3	89.1

Table 5.2: The performance results on the development sets of our greedy incremental generative system, predicting one (1w), three (3w) or five (5w) words at a time, and the ZGen system (Puduppully et al., 2016) for comparison.

to be complementary to the BLEU measure and can help us distinguish between the types of errors that a system makes.

5.3.3 Results and discussion

Table 5.2 reports the performance on the development sets of the greedy lexicalised and unlexicalised models predicting one word at a time and their variants predicting three and five words. We can see the performance of ZGen as an upper bound on the performance of our systems. The purely incremental word-by-word linearisation system constitutes the lower bound while the system predicting five words at a time is the most powerful out of the variants of our model and is the closest to a system with global search. While the 5-word prediction corresponds effectively to conducting a global search for many subtrees, we assume that the system predicting three words at a time is a cognitively plausible compromise in terms of degree of incrementality and computational load. To demonstrate that performance generalises to the test sets we report the BLEU scores of our lexicalised model predicting three words and ZGen in Table 5.3. We discuss first the results of our model in comparison to ZGen. In

	English	Italian	Persian	Russian
ZGen	84.2	81.7	79.7	67.3
Lex 3w	78.4	72.3	75.5	58.7

Table 5.3: The results on the test sets (BLEU).

the second part of this section, we look at the cross-linguistic patterns in sentence linearisation accuracies and connect them to the language-level measures of word order variability analysed in Chapter 3.

First, we can see that our basic greedy unlexicalised model shows quite low performance compared to ZGen (Table 5.2). However, much better results are obtained if we linearise predicting several words at a time. Predicting three words improves the performance by 7 to 10 BLEU points. The further improvement obtained by predicting five words is around 2 points. This small improvement over 3w model suggests that an incremental greedy prediction of only three words approximates rather well the best solution that can be found by global search. In addition, partial lexicalisation helps to improve performance, by around 0.5 (Persian) to 1.7 (Italian) points for the 3w model. Note that, compared to the lexicalised system of ZGen and our partially-lexicalised model, the unlexicalised models obtain relatively high BLEU scores suggesting that a large part of the sentence linearisation system is conditioned only on the underlying syntactic representation. Overall, the psychologically interesting lexicalised system predicting three words compares favourably to the state-of-the-art system, with BLEU scores below by 4.2 (Persian) to 9.6 (Italian) on the test sets (Table 5.3). The main reason for the inferior performance of our system compared to ZGen seem to be the strong independence assumptions that we adopt: even if we output the best global solution our model does not reach the performance of ZGen. Interestingly, our models perform much better in terms of BLEU scores than the globally-optimised models of Futrell and Gibson (2015), who report BLEU scores of only 57.7 for the English UD treebank.

The other important reason for the inferior results of our system compared to ZGen is the absence of phrase length features in the architecture of our model. Our model

does not take into account the length of dependencies and therefore cannot predict any DLM-related word order patterns, such as the fact that shorter phrases precede longer phrases in the postverbal domain. Section 5.4 presents a modified version of our model which has access to dependency length information and explicitly incorporates the choice between alternative grammatical word orders in incremental generation architecture.

Cross-linguistic analysis of the results

Across the four languages, English is the easiest to linearise while the performance numbers on Russian are significantly lower than the other three languages, despite the larger amount of the training data. Intuitively, these results should be due to English having strict word order constraints and Russian having rather flexible and discourse-conditioned word order. However, as we have argued in Chapter 3, it is hard to make certain cross-linguistic comparisons based on treebanks which differ in many properties in addition to the word order. To conduct a more informed analysis of the results across languages, we adopt the evaluation proposal of Nivre and Fang (2017) and break down the BLEU and arc direction performance numbers for different types of dependencies (Table 5.4). Nivre and Fang (2017) note that the differences in the number of words per sentence between isolating languages and morphologically-rich languages bias the performance measures to give higher parsing performance values to the languages of the first group. In our case, this could mean that worse results for Russian compared to, e.g., English or Italian, are due to the fact that these languages have easier-to-order words such as determiners (absent in Russian) or prepositions (often substituted by case marking in Russian). However, the numbers in Table 5.4 show that this confounding factor does not change the overall picture. We report the performance for a 'core' subset of dependencies including *nsubj*, *dobj*, *xcomp* and other dependencies belonging arguably to the predicate-argument structure of the sentence. All languages show lower performance for core dependencies, but the drop for English (7.3 BLEU) is smaller than for other languages, including Russian (10.2 BLEU points drop). This result suggests that core dependents in English have less word order variability than in other languages (Italian, Russian), which in turn

	BLEU		arc dir %	
Language	all	core	all	core
English	77.3	70.3	95.6	96.5
Italian	72.0	60.3	92.3	86.5
Persian	75.4	65.4	97.2	99.0
Russian	58.5	48.3	87.7	82.0

Table 5.4: Results of the unlexicalised model predicting three words broken down for all/core dependencies.

makes it a plausible reason for the low overall performance results on these languages. Note, additionally, that the linearisation systems we evaluate (including ZGen) do not have access to the morphological annotation. Russian is a case-marking language, and some of the functional relations are encoded using case suffixes. The primary distinctions between subject (nominative case), object (accusative case) and indirect object (dative case) are encoded in dependency labels, but some others are not. Finally, the Russian treebank is a collection of literary texts as opposed to newswire and internet-crawled texts of other treebanks which can be one more reason for very low sentence linearisation performance on this language.

Another interesting cross-linguistic observation concerns the arc-direction accuracy results. For ZGen and our linearisation system (both for 'core' and for all dependencies), the following ranking with respect to the arc-direction accuracy holds: Persian $>$ English $>$ Italian $>$ Russian.[10] Importantly, the arc-direction accuracies correlate with the reverse arc-direction entropy values that we computed in Chapter 4. More precisely, Persian has the smallest arc-direction entropy (0.16) and has the highest arc-direction accuracy in the linearisation (97–98%). The reverse is true for Russian: the entropy value of the Russian treebank is 0.45 and the arc-direction accuracy is only 88–91%. The fact that for Persian and English arc-direction accuracy is very high suggests that low BLEU scores are due to the variability in the order of sister phrases as opposed to the order of heads and their children. In particular, the 99%

[10]Note that the arc-direction accuracy ranking is different from the ranking based on BLEU scores.

accuracy on core dependencies in Persian reflects the fact that it is an SOV language and therefore the order of main constituents such as the subject and the object are always on the left of their head — the verb.

More generally, arc-direction accuracy can be seen as an alternative way to measure variation in the order of head and its dependents. The higher the arc-direction accuracy, the less arc-direction variation in a language. We have argued in Chapter 3 that arc-direction entropy is a robust measure of word order variation. However, its computation is subject to strong independence assumptions, e.g., that the position of a child depends only on its PoS tag and the PoS tag and the dependency label of its head. These assumptions are relaxed in the sentence linearisation systems. In our system, for example, the position of a child also depends on the previously uttered words in the phrase (n-gram probabilities).[11] Moreover, the fact that two systems with different architectures — ZGen and our linearisation system — obtain similar arc-direction accuracies suggests that they capture and indirectly quantify the same aspect of word order variation. Overall, we would like to propose that a sentence linearisation system can be used to quantify word order properties discussed in Chapter 3 such as the degree of word order variability in a language. The use of sentence linearisation systems can prove to be more robust across different languages and treebanks thanks to the generalisation ability of the underlying machine learning algorithms. In addition to arc-direction accuracy, other automatic evaluation measures adopted in the sentence linearisation task, such as BLEU score, might be used to measure new aspects of word order variation. BLEU captures, in particular, the errors in the ordering of sister phrases and it would be interesting to quantify this aspect of word order variation which is hard to do robustly (Section 3.1).

[11] In a set of preliminary experiments, we tested a linearisation model where children were placed with respect to their heads according to only the arc-direction probabilities. This model produced substantially lower results, both in terms of BLEU scores and arc-direction scores, than a model using both arc-direction and n-gram probabilities. These results confirm that arc-direction variation is conditioned on the broader context than the one used to compute arc-direction entropy in Chapter 3.

5.4 Word order variation as a re-ranking mechanism

This section presents an extension of our sentence linearisation model which incorporates explicitly the choice between two alternating word orders. Our goal is to develop a system which can efficiently condition its decisions on dependency length features and can be subsequently used to study word order variation and DLM phenomena in a new principled way, as part of a general incremental word order production process.

The basic sentence linearisation system presented in the previous sections is characterised by its generative architecture. We argued that this property is desirable for a cognitively plausible model of sentence linearisation if we assume that speakers store probabilistic grammatical knowledge and use it in comprehension and production. In fact, as part of our linearisation system, we learn a probabilistic dependency grammar defined by probabilities $P(n_0 \dots h \dots n_i \mid h)$ over expansion rules $\boldsymbol{h} \rightarrow n_0 \dots h \dots n_i$. We did not include the lengths of dependencies as factors in this probabilistic grammar. One of the reasons concerns the computational difficulty. Adding the length as a factor that should be predicted by the generative model substantially increases the probability space and leads to sparse observations in the training data. Conceptually, this type of model cannot take into account the *relative* length of phrases which we know plays an important role in DLM effects in word order variation. This is because the probabilities $P(n_1 \mid p, h)$ and $P(n_2 \mid p, h)$ for the two candidate nodes n_1 and n_2 are computed independently and can take into account only the length of the generated node itself or the length factors of the nodes in the prefix. They do not take into account the presence or absence of the other candidate node, nor its length.[12]

For these reasons, we propose a discriminative mechanism to deal with cases of word order variation and conditioning factors such as lengths of dependencies. While a generative model predicts the distribution over all words in a vocabulary given the previous context $p = n_0, \dots, n_{k-1}$, a discriminative model chooses between the subset of nodes defined by the domain D_h, e.g. $n_k \in D_h \setminus \{n_0, \dots, n_{k-1}\}$. The latter

[12] To take into account the relative influence of factors of two nodes in a generative model, we would need to introduce joint probabilities for the bigrams, i.e., $P(n_1 n_2 \mid p, h)$, which would result in severe sparsity problems.

type of model can take therefore into account information about the relation between candidate nodes.

5.4.1 Description of the advanced model

This section presents the architecture of our advanced model and gives details on how we integrate the modelling of word order variation phenomena as part of our incremental language production.

We augment our basic system with an additional re-ranking step at each ordering decision. This re-ranking which, for the moment, we apply for the two best-scored candidate nodes, can be taken to mirror traditional logistic models of word order variation. Consider a familiar example of variation in the order of postnominal dependents in English (Wasow, 2002). There are commonly two alternative orders that are grammatical and semantically equivalent, e.g., for the verb-particle construction, *a boy threw the trash out* and *a boy threw out the trash*. We have seen in Chapters 2 and 4 that the choice between these two orders is commonly modeled as logistic regression incorporating syntactic, lexical and frequency features associated with the two orders (Gries, 2003; Bresnan et al., 2007). A ranking model is a natural extension of logistic regression for modelling more than one word order variation construction (Rajkumar et al., 2016). Our re-ranking model implements a single mechanism for word order variation decisions shared across all constructions in a language. Crucially, this mechanism is incorporated in our incremental model (hence, *re*-ranking) which makes it possible to analyse word order variation and dependency length effects as part of the online linearisation procedure.

We propose the following modification to our basic linearisation model, as illustrated schematically in Figure 5.5. At the moment of choosing the next word n_k given the previously uttered words $n_0, n_1, \ldots, n_{k-1}$, we first compute the scores $Score(n_0, n_1, \ldots n_k^1), Score(n_0, n_1, \ldots n_k^2), \ldots$ for each possible node $n_k^1, n_k^2, \ldots$, as in the basic model. For each of the two best-scored nodes n_k^1 and n_k^2, we compose a set of input features associated with this node, including its PoS tag and dependency label, and, crucially, its $Score(n_0, n_1, \ldots n_k^i)$. We train a pairwise re-ranker implemented as

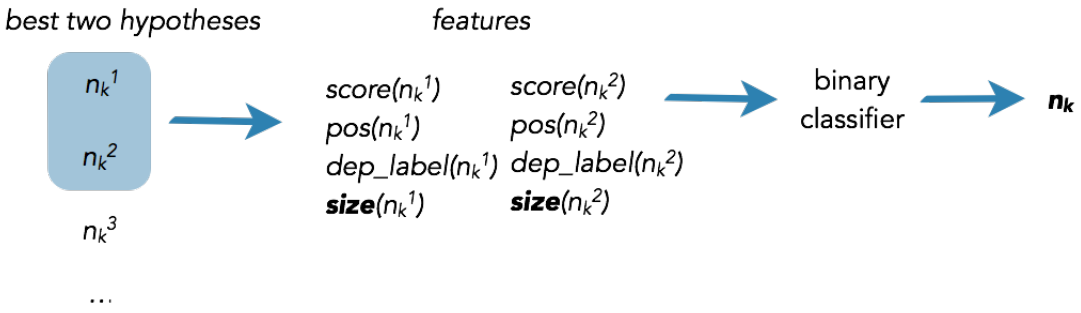


Figure 5.5: Schematic representation of the re-ranking step.

a binary classifier with perceptron update, similarly to White and Rajkumar (2009), to choose one continuation out of the two competing continuations n_k^1 and n_k^2 based on these features.

To model DLM effects associated with long phrases, we also include as an input to the re-ranker the length of the phrases headed by the nodes n_k^i, that is, the size of the dependency subtree headed by these nodes. Our discussion of the DLM principle in the previous chapters revolved around the idea that a global DLM principle generalising the observed "short-before-long" and "long-before-short" effects should take into account the lengths of dependencies as opposed to the lengths of phrases. However, the assumptions on the incremental hierarchical nature of the linearisation process that we adopted in this chapter constrain this interpretation. Indeed, to know the length of a dependency between a head and its dependent n_k^i we need to have information on the order of children in the subtree headed by n_k^i. Our recursive algorithm assumes that the linearisation (and potentially the structural expansion) of the subtrees proceed independently from each other. This assumption prevents the re-ranker to access the information about the exact length of the dependency between the head and n_k^i. In principle, we could obtain an estimation of this length by predicting the average position of a head with respect to its children. For example, in a head-initial language, we would expect the head to be close to the left edge of the phrase. For the current experiments, we take the sizes of the phrases as an approximation of the dependency lengths. We make, therefore, a weak assumption that we know the size of the subtree n_k^i (while the exact ordering of the children of n_k^i is still not known).

This discussion makes it clear that the architecture of the production system creates some complications for a production account of the global DLM principle. This is, perhaps, unsurprising given that the global DLM principle assumes minimisation of the total dependency length of the sentence and incremental production assumes instead at least some non-global decisions. These issues become much more transparent when a bias in language production is sought to be explained as part of a working implementation of the production mechanism.

Taken as a computational model of word order variation, our binary re-ranking model differs from traditional logistic word order variation models, including the ranking model of Rajkumar et al. (2016), in two aspects. First, the crucial property of our approach is that we model word order variation as part of the sentence linearisation process by including the scores provided by the underlying language production model and the acquired generative probabilistic grammar. Secondly, the choice between two alternative word orders is naturally made at the point when the two hypotheses diverge. In comparison, a typical logistic model is based on the comparison of complete alternative orders and ignores the mechanisms of language production. In our model, the preference for shorter dependencies in word order variation and at the language level should come from local decisions of the incremental linearisation system.

In this respect, our approach can also be seen as an adaptation of the work of White and Rajkumar (2009, 2012) for incremental linearisation. White and Rajkumar (2009) presented a system for re-ranking of the n-best list of complete linearisations produced by a CCG grammar generator using averaged perceptron training. White and Rajkumar (2012) enhanced their global re-ranker with dependency length features to produce more natural sentences where, e.g., shorter verb complements precede longer ones. In contrast to our approach, the work of White and Rajkumar (2009, 2012) uses global search over possible CCG derivations and applies re-ranking to the list of candidate complete sentences as the last step of the linearisation algorithm. This model, therefore, is not incremental and cannot be taken as a cognitively-plausible model of the sentence linearisation process. Our approach shows that we can use similar re-ranking techniques also as part of an incremental system.

	English		Italian		Persian		Russian	
Systems	BLEU	avg DL	BLEU	avg DL	BLEU	avg DL	BLEU	avg DL
Gold	-	2.42	-	2.45	-	3.44	-	2.59
Unlex 3w	77.3	2.57	72.0	2.76	75.4	3.55	58.5	3.08
+ reranking	74.9	2.55	70.7	2.80	75.1	3.53	58.3	2.96
+ length features	76.1	2.51	73.2	2.65	75.6	3.52	59.9	2.77

Table 5.5: Results of the model with additional re-ranking applied. The performance numbers of the basic model predicting three words are given for comparison.

5.4.2 Results and discussion

Table 5.5 reports the performance of the unlexicalised model predicting one word at a time with the binary word order variation re-ranker with and without phrase length features. For comparison, we include the performance of our basic unlexicalised model predicting three words at a time.

First of all, we see a clear improvement with an additional re-ranking step over the model predicting one word at a time with performance coming close to the reference model predicting three words. This result suggests that the information that the classifier exploits, namely the knowledge about the two best continuation nodes and their scores, is very relevant for choosing the best hypothesis. These results show that it is possible to obtain an efficient greedy model by trading in some discriminative information.

Adding phrase length features improves the performance across all four languages, by 0.5 (Persian) to 2.5 (Italian) BLEU points, confirming that lengths of phrases are significant factors affecting word order variation. Note that for Italian and Russian the resulting BLEU scores (73.2 and 59.9, respectively) are higher than the reference 3w model scores (72.0 and 58.5).

To better observe the effect of phrase lengths, we report the average dependency

lengths (DLs) for gold trees and trees linearised by our systems (Table 5.5). We find that the increase in performance goes along with the decrease in average DL. The systems aware of phrase lengths produced significantly shorter DLs compared to the reranking systems without length features and the basic unlexicalised models; the gold order is the one having the lowest DLs for all languages. Interestingly, the languages where the improvement is the smallest (Persian, 0.5 BLEU, 0.01 difference in average DL and English, 1.2 BLEU, 0.04 difference in average DL) are the ones where the average DL without length features is already close to the gold average DL. These results can be seen as a confirmation of the general DLM principle, discussed throughout this thesis, using a new computational methodology based on application and evaluation of sentence linearisation models.

Similarly to the arc-direction accuracy, the differences in the performance between the systems with and without length features can be seen as an alternative measure of the degree of DLM in a language. Intuitively, the more the system performance improves when using length features, the more the language minimises dependencies (compared to a linearisation system unaware of phrase lengths). Yet, there is one important difference with respect to our measure of DLM ratio (Section 3.1): the sentence linearisation system only produces linearisations which are consistent with the grammar it learnt during the training. It never generates completely ungrammatical orders violating local constraints such as the permutations in the examples in Figure 3.3 yielding minimal DLs. Recall that such permutations are the basis of comparison for DLM ratio computation. In this respect, the DLM measure based on sentence linearisation accuracy is related to the previous work which compared the actual DLs of sentences with the optimal-DL orders constrained by the language grammar (Gildea and Temperley, 2010; Futrell et al., 2015b). The DLM ratio measure from Chapter 3 captures both the minimisation at the level of word order variation and at the level of grammar. By contrast, the degree of DLM computed by means of the sentence linearisation evaluation reflects more the degree of DLM in word order variation since many ordering choices are fixed by the grammar. For example, Persian shows very little minimisation of dependencies according to the decrease in average DL (only 0.01), but it has very high DLM ratio value of 1.99 (Table 3.6). This apparent conflict of the two measures is cleared by the observation that Persian is an SOV

language. Because the verb is sentence-final, the dependencies between the subject and the verb are longer in an SOV language than in an SVO language. The high DLM ratio stems, therefore, from the low rate of DLM at the grammar level. The small increase in sentence linearisation performance with phrase length features suggests instead that there is a high rate of DLM at word order variation level. Experiments with sentence linearisation systems can, therefore, measure in a new systematic way the rate of DLM in word order variation, teasing it apart from inherently grammatical differences in dependency length across languages. The linearisations produced by the system with phrase length features still have longer dependencies on average compared to the gold word orders. Even if the improvement on Persian is only 0.01, the gold orders have shorter dependencies (by 0.08 on average) suggesting that a part of word order variation conditioned on dependency lengths is not captured by the system. These questions open exciting avenues for future work.

5.5 Conclusions

In this chapter, we have proposed the first step towards systematic computational modelling of word order phenomena as part of the language production process. We suggested to model word order as a sentence linearisation process — the part of the production process which incrementally maps the syntactic representation of a sentence onto the order of words. We argued that a psycholinguistically-plausible sentence linearisation model should have incremental architecture and make use of the probabilistic syntactic knowledge of speakers. The sentence linearisation model we developed is based on simple generative probabilistic grammars traditionally used as models of sentence comprehension, e.g., in statistical parsing. While a generative model naturally incorporates probabilistic lexical and syntactic knowledge of speakers, it poses challenges for efficient search of the best probable word order. Speakers produce sentences fluently without pausing for a long time, suggesting that the search for a good linearisation should be efficient. We showed that with a limited kind of lookahead a generative system could reach good performance compared to the state-of-the-art fully discriminative system with a beam search. Moreover, it does

so while keeping the hypothesis search space small and, consequently, the processing time and memory resource demands relatively low.

The re-ranking version of the linearisation system proposes a way to choose the next word based on the comparison between two possible continuations. As our results show, by trading in some discriminative information, we can use the fast greedy generation of the next word and reach performances similar to the model exploring a larger search space predicting three words at a time. Re-ranking can be seen as a way to model word order variation in a way similar to the traditional logistic models. Importantly, this architecture also allows us to condition word order choices on factors such as the lengths of phrases. Incorporating phrase length features into the model leads to better performance and smaller average DLs — a result which confirms the general DLM principle in a novel way. One of the results of our evaluation of the linearisation systems across four languages — English, Italian, Russian and Persian — is that we can use sentence linearisation systems to measure language-level word order properties such as the rate of DLM or word order freedom. These systems can be used therefore as robust (although more computationally involved) alternatives to treebank-based statistical measures that we have analysed in Chapter 3.

From a more theoretical perspective, we proposed a formal model of how DLM preferences could be operating in language production. Our proposal opens a new approach for investigations of word order variation phenomena and DLM effects. Looking further on, the integrated generative model of word order production and variation should allow us to investigate interactions between DLM, probabilistic factors and the cognitive properties of the model architecture — the incrementality and modularity. These cognitive constraints imply limited access to lexical items and imprecise estimation of the dependency lengths. If the processing system indeed exhibits these limitations, they should surface in the observed word order choices. A related question is how the processing mechanism involved in written production differs from that in spoken production. Many DLM effects were previously observed both for written and spoken data (Wasow, 2002; Bresnan et al., 2007; Francis and Michaelis, 2017). However, written texts differ from speech because they could be affected by comprehension biases as well as production biases. Contrasting computation models of comprehension (sentence processing) and production (sentence

linearisation) could lead to a theory-driven analysis of differences in written and spoken word order variation. Such experiments could be interesting to conduct, in particular, on head-final languages. The processing of, e.g., dependents of the verb, must happen without access to the head (verb) which appears later in the sentence. In comparison to a comprehension system, a production system can, presumably, condition its ordering decisions on the information given by the head.

Chapter 6

Conclusions

The work presented in this thesis examined several aspects of word order variation and dependency length minimisation. We looked at these phenomena in syntactic production as observed at three major levels of abstraction: in word order distributions over all structures and sentences in a language, in one isolated construction of word order alternation, and in the linearisation choices during online production.

These three interdisciplinary studies vary in the data they look at and the methods they use. They make connections to several independent domains of linguistic research including quantitative syntax, language processing, language production, typology and natural language processing. Ultimately, however, these studies were conducted to answer the same question: how do word order variation and dependency length minimisation work in languages? A computational solution to this problem consists in constructing a model which explains and predicts word order distribution at a given level of abstraction by incorporating properties and constraints leading to observed DLM effects. A general cross-linguistic model of word order variation is hardly at reach, but we provided several new empirical facts and methods that contribute to its development.

6.1 Contributions

Chapter 3 analysed word order and a general global version of the DLM principle at the language level. One can define DLM as a tendency to minimise the distance between all types of dependencies in a language. Given a treebank, this description can be formalised as a tendency to minimise the sum of the lengths of all annotated dependencies. This formalisation is useful from several perspectives. First, it can be used to compute an overall rate of DLM in a language (Temperley, 2008; Gildea and Temperley, 2010; Futrell et al., 2015a). We showed that the rate of DLM could be computed in such way as to allow meaningful quantitative comparison of languages with respect to this property and answer typological questions such as "Does this language minimise dependencies more than the other one?". The DLM-ratio measure enabling cross-linguistic quantitative comparisons is valuable for developing a typology of free word order languages. On the one hand, long dependencies are intuitively correlated with flexible word order. On the other hand, we found that DLM ratio and word order flexibility (measured as arc-direction entropy) capture distinct aspects of word order, as suggested by the diachronic investigation of the Latin and Ancient Greek treebanks. In future, the language-level measures of word order properties such as DLM ratio and arc-direction entropy can serve to elaborate the typology of languages of the world quantitatively. In this work, we demonstrated that these measures are also useful for NLP research, particularly, as estimates of difficulty for statistical parsing. NLP parsing systems are known to encounter difficulties in identifying long-distance dependencies (Rimell et al., 2009). They also obtain lower performance on morphologically-rich languages with flexible word order than, for example, on English (Seddah et al., 2013). The use of DLM ratio and arc-direction entropy allowed us to confirm at large-scale and across many languages the negative effect of these word order properties on parsing performance. To meaningfully compare the effect of word order on parsing performance across different languages and treebanks, we proposed a method to control for confounding factors such as the size of the training set, the average sentence length and others. The method is based on creating permuted versions of sentences to manipulate a word order property, e.g., by minimising the lengths of dependencies or by minimising the amount of word

order variation. The minimal pairs of sentences created in this way differ only in their word order but share all other characteristics. The proposed method is general and can be applied to test word order properties other than the rate of DLM and the amount of word order variation to diagnose parsing performance in depth.

The global DLM principle formulated for all dependencies annotated in a treebank is also a good starting point for analysing specific word order variation constructions. In Chapter 4, we approached in this way adjective-noun order variation in Romance languages. Our goal was to find and test potential DLM effects conditioning the choice of adjective position. Adjective-noun order variation is particularly interesting since it is structurally different from the majority of word order variation constructions which were evoked as evidence for the DLM principle. The adjective-noun construction involves the variation in the order between the modifier and the head, i.e., XP H vs H XP orders, while previously studied constructions frequently involve the variation in the order of sisters on the one side of the head, i.e., H XP YP vs H YP XP order. We showed that, from the perspective of the global DLM principle, the first type of variation affects the lengths of several dependencies, not only the adjective-noun dependency, compared to the second type of variation which only concerns two dependencies H–X and H–Y. We formulated the predictions of the DLM principle in this novel structural configuration and tested them on the data of five Romance languages: Italian, Spanish, French, Catalan and Portuguese. To our knowledge, this is the first large-scale corpus study of adjective variation involving more than one Romance language. We found effects of minimisation, consistent across all five languages, for several dependencies. The dependency between the noun and the adjective, the dependency between the noun and its postnominal dependent (Y) and the dependency between the noun and its head (X) are all minimised but to a different extent. The minimisation of the N–X and, especially, the N–Y dependencies is a new empirical fact for the syntactic puzzle of the adjective-noun variation. These data also highlight the existence of diverse types of DLM effects. For instance, that the minimisation of N–Y dependency is affected by the type of lexico-semantic relation between the noun and its right dependents (a PP or a relative clause).

Chapter 5 provides a new perspective on computational modelling of word order variation and DLM. DLM effects are commonly observed in spoken and written

production data. What is the mechanism in production that is responsible for these effects? If it is a processing pressure, how is it incorporated in the general language production process and how does it interact with word order planning? We suggested that we can provide initial answers to these questions by developing a psychologically plausible model of sentence linearisation — one module of the overall language production process which maps the hierarchical syntactic representation onto the order of words. We argued that cognitive plausibility means at least two things: that the order is constructed incrementally and that the choice of the next word is based on probabilistic grammatical knowledge of speakers. These requirements arise from the experimental evidence on language production. For instance, we know that speakers do not plan entire sentences before they start speaking which means that word order cannot be optimised globally and there should be some degree of online processing in deciding which word or phrase to say next. Our model is based on incremental recursive linearisation of dependency subtrees. The next word to produce (out of the nodes in a subtree including the head and its children) is chosen using a score function. Our probabilistic score function combines generative syntactic probabilities and limited lookahead in the form of a future score to provide accurate but, to a large extent, local decisions. The resulting linearisation process is incremental since the model chooses one or several next words at each step greedily. We evaluated the model on four languages and compared it to the state-of-the-art sentence linearisation statistical model without incremental constraints. We found that the incremental model achieved relatively good performance compared to the state-of-the-art confirming the empirical plausibility of the imposed architectural constraints. In addition, we experimented with a mechanism to incorporate the choice between alternative word orders in the incremental architecture. This choice was conditioned on features related to dependency lengths. The variation and DLM aspects of word order were naturally incorporated in a discriminative re-ranking step on top of the score function decisions of the basic linearisation model. We showed that conditioning the choice between two possible continuations on the phrase length features improved performance of the models. The fact that the overall dependency length of the output trees decreased with the improved performance confirms once again that languages prefer orders which minimise dependencies. Our study suggests, more generally, that incremental sentence linearisation models are a

promising tool in studying word order variation phenomena from the perspective of language production and that they can be used as a more robust alternative to quantify language-level word order properties such as the rate of DLM or arc-direction entropy.

6.2 Future work

Three computational treebank-based approaches to word order and DLM presented in Chapters 3 to 5 open a lot of interesting possibilities for future research. We already discussed some ideas for direct continuations of these studies at the end of each chapter. In this section, we would like to briefly address the general emerging picture on DLM as well as some related topics that could not make part of this thesis.

6.2.1 Unified account of processing-related biases

DLM in its most general formulation can refer to a large number of effects in language production, ranging from typological tendencies across languages to categorial constituency rules in language grammars to fine-grained gradual preferences in word order alternation constructions. DLM effects in word order variation can be observed both in long and short spans (Chapter 4). It is tempting to provide an account of all these seemingly related phenomena in terms of one universal principle. Such principle could stem from communication efficiency pressures (including processing load) shaping language through its evolution. This is a view advocated by Hawkins (2004) and, very recently, by Futrell and Levy (2017) who propose a new general principle of "information locality", subsuming the DLM principle.

A straightforward and intuitive way to extend the DLM principle defined based on the syntactic notion of dependencies is to condition the "strength" of minimisation on the type of dependency. As mentioned in Chapters 2 and 4, Hawkins (2001, 2004) proposes the Minimisation of Domains (MiD) principle (an extension of his purely syntactic Early Immediate Constituents principle) as a formalisation of the intuition

that semantically-involved dependencies should be minimised more strongly than purely syntactic dependencies. Lohse et al. (2004) adopt MiD principle to explain the variation in the verb-particle split construction in English.[1] They analyse the minimisation effects between verbs and particles using two semantic dependencies defined in terms of entailment. First, there is a verb-particle dependency if the verb meaning is dependent on the particle, that is, the meaning of the entire verb+particle phrase (e.g., *turn off the lights*) does not entail the meaning of the verb in isolation (*turn*). Secondly, there is a particle-verb dependency if the particle meaning is dependent on the verb, that is, the meaning of the verb+particle phrase (e.g., *call parents up*) does not entail the meaning *be/become/going*+particle (e.g., *going up*). Unfortunately, this definition is problematic for many reasons, including the polysemy of verbs and particles and the reliance on entailment judgments provided by human annotators. Consequently, it cannot be operationalised on a large scale and applied in the same way to all types of dependencies.

An alternative simple and intuitive measure of "strength" of a dependency is pointwise mutual information (PMI) between the head and the dependent $PMI = \log \frac{p(h|d)p(d|h)}{p(h,d)}$. The definition of PMI relies on the probabilities of observing the head $p(h)$ and the dependent $p(d)$ and their co-occurrence $p(h,d)$. High PMI means that h and d often co-occur, in other words, the presence of h is highly predictive of d and vice versa. PMI is symmetric and does not take into account the semantics of the items. It can be estimated, for example, from co-occurrence statistics of two lexical items in a large enough corpus. Following the experiments in Section 4.3, we conducted preliminary investigations of adjective variation in Italian to test whether PMI affects the minimisation of the dependency between the noun and its right dependents (such as prepositional phrases) (Gulordava, 2016). We found a significant effect of PMI but also of other related — and correlated — factors such as the frequency of the N-PP phrase and the conditional probability $p(PP|N)$. The use of probabilistic measures such as PMI is particularly interesting since it can be related to the notion of surprisal (Hale, 2001; Levy, 2008; Rajkumar et al., 2016). In comprehension, high surprisal values explain difficulties observed in sentence processing at the moment when an

[1] Wiechmann and Lohmann (2013) apply MiD principle in a similar way to analyse alternation in the order of prepositional phrases in postveral domain.

unexpected word must be processed and incorporated in a partially constructed syntactic structure. Conditional log-probability $\log p(PP|N)$ is the surprisal of the prepositional phrase given the noun.

Futrell and Levy (2017) propose that DLM is derived from a more general information locality principle under the assumption that dependencies connect pairs of words that have higher mutual information than other pairs of words. Information locality is proposed as a general communication efficiency principle — a consequence of the fact that language is a noisy-channel communication tool. The assumption that noise happens during transmission, e.g., in the form of deletion errors, leads to the conclusion that elements which are predictive of a word have to be close linearly to facilitate the processing and interpretation of this word. Futrell shows that words in a head-dependent relationship are on average more predictive of each other, i.e., have higher PMI than other pairs of elements in the sentence[2] Consequently, the information locality principle should apply for the dependent elements, leading to what we call the DLM principle.

The information locality principle, inducing formally the idea that PMI could be taken as a measure of dependency "strength", provides an interesting high-level account of DLM. However, it is not clear to what extent the DLM effects observed in specific constructions and word order variation phenomena can be explained using this new tool. Indeed, if languages respected the "higher PMI — more adjacent position" principle rigidly, unsupervised syntax induction would be an easy task, which it is not (Klein, 2005). Our preliminary results on adjective variation also suggest that the simple PoS-tag based PMI measures cannot account for fine-grained preferences between Adj N PP and N Adj PP orders. An extensive cross-linguistic analysis of PMI and related probabilistic factors is required to establish better the overall picture. The use of incremental generative linearisation models, developed in Chapter 5, can also prove useful to test PMI-related predictions. Our linearisation model explicitly incorporates probabilities which means that PMI or conditional surprisal values can be used as factors conditioning word order and dependency lengths during

[2]These results are based on dependencies defined by UD annotation and statistics on co-occurrence of PoS tags of pair of elements.

linearisation. Such a model can be used to test systematically and cross-linguistically the general effect of these factors across all types of dependencies.

6.2.2 Non-projective order and DLM

In this thesis, we did not touch on the question of non-projective dependencies and their relation to DLM. In particular, our sentence linearisation model (Chapter 5) can produce only projective structures. While some natural languages exhibit a significant degree of non-projectivity its occurrence is limited in the four Indo-European languages which we used in our linearisation experiments: there is only 0.2-0.3% of non-projective arcs in English, Italian and Persian and 0.7% in Russian. These percentages are relatively small because the flat content-head dependency annotation of UD favours low non-projectivity.

Also, we have not discussed the effect of non-projective dependencies on parsing analysis in Chapter 3. The challenge here is that the presence of long dependencies is strongly correlated with non-projectivity, as predicted theoretically (Ferrer-i-Cancho, 2006). The two properties — DLM ratio and percentage of non-projective dependencies — measured on our UD treebank data have the Pearson correlation of 0.66. While this correlation is strong, it is not perfect and suggests that both dependency length and non-projectivity should be taken into account to explain parsing performance values.

Despite a strong empirical and theoretical association between non-projective structures and long dependencies, interestingly, non-projectivity sometimes leads to word orders which minimise dependency length. Such cases include, for instance, extraposition constructions in English, exemplified in (6.1) (Levy, 2005; Francis, 2010).

(6.1) a. Evidence [that shows a new side effect of the medicine] has been found.
b. Evidence has been found [that shows a new side effect of the medicine].

In this example, there are two dependencies of interest: the dependency between *evidence* and the extracted relative clause *that shows a new side effect of the medicine* and

the dependency between the verb *found* and the subject *evidence*. It is not difficult to see that the second order, which involves extraposition, is the one with the shorter overall dependency length. Yet, out of the two structures, the non-projective extraposition is more difficult to process, and this difficulty correlates with the surprisal measure of processing load (Levy et al., 2012). Interestingly, there have been found differences between comprehension and production of the extraposed constructions (Francis, 2010; Francis and Michaelis, 2017). In a recent study, Francis and Michaelis (2017) conduct and compare a preference experiment and an elicited production experiment of extraposition constructions. They show, among other results, that the relative effect of the length of the two manipulated phrases (VP *has been found* and the relative clause) is different in the two tasks.

It is an intriguing question how non-projective structures arise in production and how this mechanism can be implemented in a linearisation model, especially since they seem to violate the assumption of incremental hierarchical planning. An apparent violation arises because the VP and the relative clause under the noun head belong to different subtrees. One explanation that was proposed is the "easy-first" principle (MacDonald, 2013), which suggests that simpler and shorter to produce phrases are spelt out first. Such principle could explain the extraposition construction and "short-before-long" preferences, but it is not clear how it applies to the "long-before-short" effects in head-final languages and how it should be exactly incorporated in the incremental planning (Jaeger and Norcliffe, 2009).

There exist a puzzling interaction of different processing pressures in non-projective extraposition constructions. Further investigation and computational modelling are necessary to shed light on how these pressures operate and differ in comprehension and production.

6.3 Conclusions

Word order is, perhaps, the most fundamental observable property of the grammar. It is hard to understate the importance of the studies on word order in linguistics and the number of questions that are still unresolved. Variation and DLM phenomena

in word order are two major topics which cut through several domains of language study: syntax, production and comprehension. We believe this thesis can be taken as a testimony to the fruitful approach of combining tools and ideas from several linguistic research traditions and looking at the same questions from various points of views. We are optimistic that this perspective and the results of the thesis will inspire further contributions to the continuous study of language and word order.

Bibliography

Anne Abeillé and Danièle Godard. La position de l'adjectif épithète en français: le poids des mots. *Recherches linguistiques de Vincennes*, 28:9–32, 1999.

Anne Abeillé and Danièle Godard. French word order and lexical weight. In Robert D. Borsley, editor, *The nature and function of Syntactic Categories*, volume 32 of *Syntax and Semantics*, pages 325–360. BRILL, 2000.

Anne Abeillé and Danièle Godard. The syntax of French adverbs without functional projections. *Current studies in comparative Romance linguistics. Amsterdam: John Benjamins*, pages 1–39, 2003.

Željko Agić, Maria Jesus Aranzabe, Aitziber Atutxa, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Jan Hajič, Anders Trærup Johannsen, Jenna Kanerva, Juha Kuokkala, Veronika Laippala, Alessandro Lenci, Krister Lindén, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Héctor Alonso Martínez, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Joakim Nivre, Hanna Nurmi, Petya Osenova, Slav Petrov, Jussi Piitulainen, Barbara Plank, Prokopis Prokopidis, Sampo Pyysalo, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Kiril Simov, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. Universal dependencies 1.1, 2015. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Artemis Alexiadou. Adjective syntax and noun raising: word order asymmetries in the DP as the result of adjective distribution. *Studia linguistica*, 55(3):217–248, 2001.

Artemis Alexiadou. *Adverb placement: A case study in antisymmetric syntax*, volume 18. John Benjamins Publishing, 1997.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally Normalized Transition-Based Neural Networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2442–2452. Association for Computational Linguistics, 2016.

Jennifer E. Arnold, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. Heaviness vs. Newness: The Effects of Structural Complexity and Discourse Status on Constituent Ordering. *Language*, 76(1):28–55, 2000.

Harald R. Baayen, Douglas J. Davidson, and Douglas M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412, 2008.

Miguel Ballesteros and Joakim Nivre. MaltOptimizer: An optimization tool for MaltParser. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*, pages 58–62, Avignon, France, April 2012.

David Bamman and Gregory R. Crane. Building a dynamic lexicon from a digital library. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (JCDL'08)*, pages 11–20, New York, NY, USA, 2008. ACM.

David Bamman and Gregory R. Crane. The Ancient Greek and Latin Dependency Treebanks. In Caroline Sporleder, Antal Bosch, and Kalliopi Zervanou, editors, *Language Technology for Cultural Heritage*, Theory and Applications of Natural Language Processing, pages 79–98. Springer Berlin Heidelberg, 2011.

Srinivas Bangalore and Owen Rambow. Exploiting a probabilistic hierarchical model for generation. *Proceedings of the 18th conference on Computational linguistics -*, 1: 42–48, 2000.

Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. *lme4: Linear mixed-effects models using Eigen and S4*, 2014. R package version 1.1-7.

Otto Behaghel. *Deutsche Syntax: eine geschichtliche Darstellung. Band IV: Wortstellung-Periodenbau*. Carl Winter, Heidelberg, 1932.

Anja Belz, Michael White, Dominic Espinosa, Eric Kow, Deirdre Hogan, and Amanda Stent. The First Surface Realisation Shared Task: Overview and Evaluation Results. *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG'11)*, 2(September):217–226, 2011.

Emily M. Bender, Dan Flickinger, Stephan Oepen, and Yi Zhang. Parser Evaluation over Local and Non-Local Deep Dependencies in a Large Corpus. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 397–408. Association for Computational Linguistics, 2011.

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

Lars M. Blöhdorn. *Postmodifying Attributive Adjectives in English: An Integrated Corpus-based Approach*, volume 7 of *English corpus linguistics*. Peter Lang, 2008.

Kathryn Bock and J. Cooper Cutting. Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, 31(1):99–127, 1992.

Kathryn Bock and Willem J. Levelt. Language Production. In M.A. Gernsbacher, editor, *Handbook of Psycholinguistics*, chapter 29, pages 741–779. Academic Press, New York, 1994.

Kathryn Bock, Helga Loebell, and Randal Morey. From conceptual roles to structural relations: Bridging the syntactic cleft. *Psychological review*, 99(1):150, 1992.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. The Prague dependency treebank. In *Treebanks*, pages 103–127. Springer, 2003.

Bernd Bohnet, Leo Wanner, Simon Mille, and Alicia Burga. Broad coverage multilingual deep sentence generation with a stochastic multi-level realizer. In *Proceedings*

of the 23rd International Conference on Computational Linguistics, COLING '10, pages 98–106, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Bernd Bohnet, Anders Björkelund, Jonas Kuhn, Wolfgang Seeker, and Sina Zarrieß. Generating non-projective word order in statistical linearization. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 928–939, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Marisa Ferrara Boston, John T. Hale, Shravan Vasishth, and Reinhold Kliegl. Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, 26(3):301–349, 2011.

Denis Bouchard. The distribution and interpretation of adjectives in French: A consequence of Bare Phrase Structure. *Probus*, 10(2):139–184, 1998.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. Predicting the dative alternation. In G. Boume, I. Kraemer, and J. Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Science, Amsterdam, 2007.

Sabine Buchholz and Erwin Marsi. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 149–164, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Aoife Cahill. Correlating human and automatic evaluation of a german surface realiser. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 97–100, Suntec, Singapore, August 2009. Association for Computational Linguistics.

Alberto Centeno-Pulido. *Reconciling generativist and functionalist approaches on adjectival position in Spanish*. PhD thesis, UGA, 2010.

Franklin Chang. Symbolically speaking: A connectionist model of sentence production. *Cognitive science*, 26(5):609–651, 2002.

Franklin Chang. Learning to order words: A connectionist model of heavy NP shift and accessibility effects in japanese and english. *Journal of Memory and Language*, 61 (3):374–397, 2009.

Franklin Chang, Gary S. Dell, and Kathryn Bock. Becoming syntactic. *Psychological review*, 113(2):234, 2006.

Danqi Chen and Christopher Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October 2014. Association for Computational Linguistics.

Hye-Won Choi. Length and Order: A Corpus Study of Korean Dative-Accusative Construction. *Discourse and Cognition*, 14(3):207–227, 2007.

François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

Noam Chomsky. *Syntactic Structures*. The Hague/Paris: Mouton, 1957.

Noam Chomsky. *Aspects of the Theory of Syntax*. MIT press, 1965.

Noam Chomsky. *Lectures on Government and Binding: The Pisa Lectures*. Folis Publications Holland, 1981.

Noam Chomsky. *The Minimalist Program*. MIT Press, 1995.

Noam Chomsky and Howard Lasnik. The theory of principles and parameters. *Syntax: An international handbook of contemporary research*, 1:506–569, 1993.

Fan R. K. Chung. On optimal linear arrangements of trees. *Computers & Mathematics with Applications*, 10(1):43–60, 1984.

Guglielmo Cinque. Deriving Greenberg's Universal 20 and its exceptions. *Linguistic Inquiry*, 36(3):315–332, 2005.

Guglielmo Cinque. *The Syntax of Adjectives: A Comparative Study*. MIT Press, 2010.

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(Aug):2493–2537, 2011.

Marie-Catherine de Marneffe and Christopher D. Manning. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, CrossParser '08, pages 1–8, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*, pages 4585–4592, Reykjavik, Iceland, May 2014.

Gary S. Dell, Franklin Chang, and Zenzi M Griffin. Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science*, 23(4): 517–542, 1999.

Vera Demberg and Frank Keller. A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of the Cognitive Science Society*, volume 31, 2009a.

Vera Demberg and Frank Keller. A computational model of prediction in human parsing: Unifying locality and surprisal effects. In Niels Taatgen and Hedderik van Rijn, editors, *Proceedings of the 29th meeting of the Cognitive Science Society (CogSci-09)*, pages 1888–1893, Amsterdam, 2009b. Cognitive Science Society.

David R. Dowty. Toward a minimalist theory of syntactic structure. *Discontinuous Constituency*, 6:11, 1996.

M. S. Dryer and Martin Haspelmath, editors. *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich, 2011.

Matthew S. Dryer. The greenbergian word order correlations. *Language*, 68:81–138, 1992.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209, San Diego, California, June 2016. Association for Computational Linguistics.

Jason Eisner. Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pages 340–345, Copenhagen, August 1996.

Yehuda Falk. Constituency, word order, and phrase structure rules. *Linguistic Analysis*, 11(331):60, 1983.

Fernanda Ferreira. Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, 30(2):210–233, 1991.

Fernanda Ferreira and Paul E. Engelhardt. Syntax and Production. In *Handbook of Psycholinguistics (Second Edition)*, chapter 3, pages 61–91. Academic Press, 2006.

Fernanda Ferreira and Benjamin Swets. How Incremental Is Language Production? Evidence from the Production of Utterances Requiring the Computation of Arithmetic Sums. *Journal of Memory and Language*, 46(1):57–84, 2002.

Ramon Ferrer-i-Cancho. Euclidean distance between syntactically linked words. *Physical Review E*, 70(5), 2004.

Ramon Ferrer-i-Cancho. Why do syntactic links not cross? *EPL (Europhysics Letters)*, 76(6):1228, 2006.

Ramon Ferrer-i-Cancho. Hubiness, length, crossings and their relationships in dependency trees. *Glottometrics*, 25:1–21, 2013.

Ramon Ferrer-i-Cancho and Haitao Liu. The risks of mixing dependency lengths from sequences of different length. *Glottotheory*, 5(2):143–155, 2014.

Katja Filippova and Michael Strube. Tree linearization in english: Improving language model based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational*

Linguistics, Companion Volume: Short Papers, pages 225–228, Boulder, Colorado, June 2009. Association for Computational Linguistics.

Mats Forsgren. *La place de l'adjectif épithète en français contemporain, étude quantitative et sémantique.* Almqvist & Wilksell, Stockholm, 1978.

Gwendoline Fox and Juliette Thuilier. Predicting the Position of Attributive Adjectives in the French NP. In Daniel Lassiter and Marija Slavkovik, editors, *New Directions in Logic, Language and Computation*, Lecture Notes in Computer Science, pages 1–15. Springer, April 2012.

Jean E. Fox Tree and Herbert H. Clark. Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, 62(2):151 – 167, 1997.

Elaine J. Francis. Grammatical weight and relative clause extraposition in English. *Cognitive Linguistics*, 21(1):35–74, 2010.

Elaine J. Francis and Laura A. Michaelis. When relative clause extraposition is the right choice, it's easier. *Language and Cognition*, 9(2):332–370, 2017.

Victoria A. Fromkin. The non-anomalous nature of anomalous utterances. *Language*, 47:27–52, 1971.

Richard Futrell and Edward Gibson. Experiments with generative models for dependency tree linearization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1978–1983, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Richard Futrell and Roger Levy. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 688–698, 2017.

Richard Futrell, Kyle Mahowald, and Edward Gibson. Quantifying Word Order Freedom in Dependency Corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, Uppsala, Sweden, August 2015a. Uppsala University, Uppsala, Sweden.

Richard Futrell, Kyle Mahowald, and Edward Gibson. Large-Scale Evidence of Dependency Length Minimization in 37 Languages. *Proceedings of the National Academy of Sciences of the United States of America*, 2015b.

Merrill F. Garrett. The limits of accommodation: Arguments for independent processing levels in sentence production. In Victoria A. Fromkin, editor, *Errors in linguistic performance: Slips of the tongue, ear, pen and hand*, pages 263–271. Academic Press, New York, 1980.

Merrill F. Garrett. Processes in language production. In Frederic J. Newmeyer, editor, *Linguistics: The Cambridge survey, Vol. 3. Language: Psychological and biological aspects*. Cambridge University Press, New York, 1988.

Gerald Gazdar, Ewan Klein, Geoffrey Pullum, and Ivan Sag. *Generalized Phrase Structure Grammar*. Harvard University Press, 1985.

Edward Gibson. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76, 1998.

Edward Gibson. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, pages 95–126, 2000.

Daniel Gildea and David Temperley. Optimizing Grammars for Minimum Dependency Length. In *Proceedings of the 45th Annual Conference of the Association for Computational Linguistics (ACL'07)*, pages 184–191, Prague, Czech Republic, 2007.

Daniel Gildea and David Temperley. Do Grammars Minimize Dependency Length? *Cognitive Science*, 34(2):286–310, 2010.

Yoav Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017.

Joseph H. Greenberg. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press, 1963.

Maurice Grevisse and André Goosse. *Le bon usage*. De Boeck Duculot, 14th edition, 2007.

Stefan Gries. A multifactorial analysis of syntactic variation: particle movement revisited. *Journal of Quantitative Linguistics*, 8(1):33–50, 2001.

Stefan Thomas Gries. *Multifactorial analysis in corpus linguistics: A study of particle placement*. A&C Black, 2003.

Kristina Gulordava. Lexico-semantic factors in the variation of adjective placement in complex noun phrases in italian. Distributional Semantics and Linguistic Theory (DSALT) Workshop, Bolzano, 2016.

Kristina Gulordava and Paola Merlo. Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of Latin and Ancient Greek. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 121–130, Uppsala, Sweden, August 2015a.

Kristina Gulordava and Paola Merlo. Structural and lexical factors in adjective placement in complex noun phrases across romance languages. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 247–257, Beijing, China, July 2015b. Association for Computational Linguistics.

Kristina Gulordava and Paola Merlo. Multi-lingual Dependency Parsing Evaluation: a Large-scale Analysis of Word Order Properties using Artificial Data. *Transactions of the Association for Computational Linguistics*, 4:343–356, 2016.

Kristina Gulordava, Paola Merlo, and Benoit Crabbé. Dependency length minimisation effects in short spans: a large-scale analysis of adjective placement in complex noun phrases. 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 477–482, Beijing, China, July 2015.

Yuqing Guo, Deirdre Hogan, and Josef Van Genabith. DCU* at Generation Challenges 2011 Surface Realisation track. In *Proceedings of the 13th European workshop on natural language generation*, pages 227–229. Association for Computational Linguistics, 2011.

Eva Hajičová. Prague dependency treebank: From analytic to tectogrammatical annotations. *Proceedings of 2nd TST, Brno, Springer-Verlag Berlin Heidelberg New York*, pages 45–50, 1998.

John Hale. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001.

Ken Hale. Walpiri and the grammar of non-configurational languages. *Natural Language & Linguistic Theory*, 1(1):5–47, 1983.

Dag T. T. Haug and Marius L. Jøhndal. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proceedings of the 2nd Workshop on Language Technology for Cultural Heritage Data*, pages 27–34, Marrakech, Morocco, 2008.

John A. Hawkins. Why are categories adjacent? *Journal of Linguistics*, 37(1):1–34, 2001.

John A. Hawkins. *Word order universals*. New York: Academic Press., 1983.

John A. Hawkins. *A performance theory of order and constituency*. Cambridge University Press, Cambridge, 1994.

John A. Hawkins. *Efficiency and Complexity in Grammars*. Oxford linguistics. Oxford University Press, Oxford, UK, 2004.

John A. Hawkins. The relative order of prepositional phrases in english: Going beyond manner–place–time. *Language variation and change*, 11(3):231–266, 1999.

James Henderson. Generative Versus Discriminative Models for Statistical Left-Corner Parsing. In *Proceedings of 8th International Workshop on Parsing Technologies (IWPT 2003)*, pages 115–126, 2003.

Grover Hudson. Is deep structure linear? In J.M. Meisel and M.D. Pam, editors, *Linear Order and Generative Theory*, Current Issues in Linguistic Theory. John Benjamins Publishing Company, 1979.

Richard A. Hudson. *Word grammar*. Blackwell Oxford, 1984.

Ray S. Jackendoff. *Semantic interpretation in generative grammar*. MIT Press, Cambridge, MA, 1972.

T. Florian Jaeger. Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of memory and language*, 59(4):434–446, 2008.

T. Florian Jaeger and Elisabeth J. Norcliffe. The Cross-linguistic Study of Sentence Production. *Language and Linguistics Compass*, 3:866–887, 2009.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling, 2016.

Dan Jurafsky. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Rens Bod, Jennifer Hay, and Stefanie Jannedy, editors, *Probabilistic linguistics*. MIT Press Cambridge, MA, 2003.

Richard S. Kayne. *The antisymmetry of syntax*. Number 25 in Linguistic Inquiry Monograph. MIT Press, 1994.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Dan Klein. *The unsupervised learning of natural language structure*. PhD thesis, Stanford University, Stanford, CA, USA, 2005. AAI3162386.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. *Dependency Parsing*. Morgan and Claypool, 2009.

Christopher Laenzlinger. French adjective ordering: perspectives on DP-internal movement types. *Lingua*, 115(5):645–689, 2005.

Irene Langkilde and Kevin Knight. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 704–710, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

John Lee, Jason Naradowsky, and David A. Smith. A Discriminative Model for Joint Morphological Disambiguation and Dependency Parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language*

Technologies, pages 885–894, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Philip Leifeld. texreg: Conversion of statistical model output in R to LaTeX and HTML tables. *Journal of Statistical Software*, 55(8):1–24, 2013.

Willem J. M. Levelt. *Speaking: From intention to articulation*. MIT Press, Cambridge, MA, 1989.

Roger Levy. *Probabilistic Models of Word Order and Syntactic Discontinuity*. PhD thesis, Stanford University, 2005.

Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008.

Roger Levy, Florencia Reali, and Thomas L. Griffiths. Modeling the effects of memory on human online sentence processing with particle filters. In *Advances in neural information processing systems*, pages 937–944, 2009.

Roger Levy, Evelina Fedorenko, Mara Breen, and Edward Gibson. The processing of extraposed structures in English. *Cognition*, 122(1):12–36, January 2012.

Haitao Liu. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191, 2008.

Haitao Liu. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578, 2010.

Yijia Liu, Yue Zhang, Wanxiang Che, and Bing Qin. Transition-based syntactic linearization. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 113–122, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

Barbara Lohse, John A. Hawkins, and Thomas Wasow. Domain minimization in english verb-particle constructions. *Language*, pages 238–261, 2004.

Maryellen C. MacDonald. How language production shapes language form and comprehension. *Frontiers in psychology*, 4(April):226, 2013.

Christopher D. Manning. Probabilistic syntax. In Rens Bod, Jennifer Hay, and Stefanie Jannedy, editors, *Probabilistic linguistics*, pages 289–341. MIT Press Cambridge, MA, 2003.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The penn treebank: annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology*, pages 114–119. Association for Computational Linguistics, 1994.

Ryan McDonald and Joakim Nivre. Analyzing and Integrating Dependency Parsers. *Computational Linguistics*, 37(1):197–230, 2011.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL'06)*, pages 216–220. Association for Computational Linguistics, 2006.

Ryan McDonald, Slav Petrov, and Keith Hall. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK, July 2011.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. Universal dependency annotation for multilingual parsing. In *ACL (2)*, pages 92–97, 2013.

Igor Aleksandrovič Mel'čuk. *Dependency syntax: theory and practice*. SUNY press, 1988.

Paola Merlo. Evaluation of two-level dependency representations of argument structure in long-distance dependencies. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 221–230, Uppsala, Sweden, August 2015.

Antje S. Meyer. Lexical access in phrase and sentence production: Results from picture-word interference experiments. *Journal of memory and Language*, 35:477–496, 1996.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3, 2010.

Tahira Naseem, Regina Barzilay, and Amir Globerson. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 629–637, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Joakim Nivre. Towards a universal grammar for natural language processing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer, 2015.

Joakim Nivre and Chiao-Ting Fang. Universal dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95, 2017.

Joakim Nivre, Johan Hall, and Jens Nilsson. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 2216–2219, Genova, Italy, May 2006.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. The CoNLL 2007 Shared Task on Dependency Parsing. *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, 2007.

Joakim Nivre, Laura Rimell, Ryan McDonald, and Carlos Gómez-Rodríguez. Evaluation of dependency parsers on unbounded dependencies. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 833–841, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia

Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC '16)*, Portoroz, Slovenia, May 2016.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Aljoscha Burchardt, Marie Candito, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Silvie Cinková, Çağrı Çöltekin, Miriam Connor, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Marhaba Eli, Ali Elkahky, Tomaž Erjavec, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Phuong Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Pinkey Nainwani, Anna Nedoluzhko, Luong Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Mar-

tin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Livy Real, Siva Reddy, Georg Rehm, Larissa Rinaldi, Laura Rituma, Rudolf Rosa, Davide Rovati, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Antonio Stella, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zhuoran Yu, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. Universal dependencies 2.0 – CoNLL 2017 shared task development and test data, 2017. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Timothy Osborne. Diagnostics for constituents: Dependency, constituency, and the status of function words. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 251–260, Uppsala, Sweden, August 2015. Uppsala University, Uppsala, Sweden.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

Albert Y. Park and Roger Levy. Minimal-length linearizations for mildly context-sensitive dependency trees. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'09)*, pages 335–343, 2009.

Slav Petrov, Dipanjan Das, and Ryan T. McDonald. A Universal Part-of-Speech Tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey, 2012.

Martin J. Pickering and Simon Garrod. An integrated theory of language production and comprehension. *The Behavioral and brain sciences*, 36(4):329–347, 2013.

Martin J. Pickering and Simon Garrod. Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11(3):105–110, 2007.

José C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

Carl Pollard and Ivan A. Sag. *Head-driven phrase structure grammar*. University of Chicago Press, 1994.

Ratish Puduppully, Yue Zhang, and Manish Shrivastava. Transition-based syntactic linearization with lookahead features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 488–493, San Diego, California, June 2016. Association for Computational Linguistics.

Geoffrey K. Pullum. Free word order and phrase structure rules. In James Pustejovsky and Peter Sells, editors, *Proceedings of the Twelfth Annual Meeting of the North Eastern Linguistic Society*, pages 209–220. Graduate Linguistics Student Association, University of Massachusetts, 1982.

Hugo Quené and Huub Van den Bergh. Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, 59 (4):413–425, 2008.

Rajakrishnan Rajkumar, Marten van Schijndel, Michael White, and William Schuler. Investigating Locality Effects and Surprisal in Written English Syntactic Choice Phenomena. *Cognition*, 155:204–232, 2016.

Owen Rambow, Bonnie Dorr, David Farwell, Rebecca Green, Nizar Habash, Stephen Helmreich, Eduard Hovy, Lory Levin, Keith J. Miller, Teruko Mitamura, Florence Reeder, and Advaith Siddharthan. Parallel Syntactic Annotation of Multiple Languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA), 2006.

Mohammad Sadegh Rasooli, Manouchehr Kouhestani, and Amirsaeid Moloodi. Development of a persian syntactic dependency treebank. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 306–314, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

Mike Reape. *A formal theory of word order: A case study in West Germanic*. PhD thesis, University of Edinburgh, Edinburgh, UK, 1993.

Mike Reape. Domain union and word order variation in german. *German in head-driven phrase structure grammar*, 46:151–197, 1994.

Ehud Reiter and Anja Belz. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics*, 35(4):529–558, 2009.

Jan Rijkhoff. Order in the noun phrase of the languages of Europe. In Anna Siewierska, editor, *Constituent Order in the Languages of Europe*, volume 20-1 of *Empirical Approaches to Language Typology*, pages 321–382. Mouton de Gruyter, 1998.

Laura Rimell, Stephen Clark, and Mark Steedman. Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 813–821, Singapore, August 2009. Association for Computational Linguistics.

Luigi Rizzi. *The structure of CP and IP: The cartography of syntactic structures*, volume 2. Oxford University Press, 2004.

Idoia Ros, Mikel Santesteban, Kumiko Fukumura, and Itziar Laka. Aiming at shorter dependencies: the role of agreement morphology. *Language, Cognition and Neuroscience*, 30(9):1156–1174, 2015.

Rudolf Rosa and Zdenek Zabokrtsky. Klcpos3 - a language similarity measure for delexicalized parser transfer. 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 243–249, Beijing, China, July 2015.

J Robert Ross. *Constraints on variables in syntax*. PhD thesis, MIT, 1967.

Pierre Joseph André Roubaud. *Nouveaux Synonimes François*. Moutard, 1786.

Djamé Seddah, Reut Tsarfaty, and Jennifer Foster, editors. *SPMRL '11: Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*. Association for Computational Linguistics, Stroudsburg, PA, USA, 2011.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

Petr Sgall, Eva Hajicová, and Jarmila Panevová. *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media, 1986.

Michelle Sheehan. Explaining the Final-over-Final Constraint: Formal and functional approaches. In Theresa Biberauer, Anders Holmberg, Ian Roberts, and Michelle Sheehan, editors, *The Final-over-Final Constraint: A Word-Order Universal and its Implications for Linguistic Theory*. MIT Press, Cambridge, Mass, 2012.

Michelle Sheehan. The Final-over-Final Condition and the Head-Final Filter. In Michelle Sheehan, Teresa Biberauer, Anders Holmberg, and Ian Roberts, editors, *To appear in The Final-over-Final Condition*. MIT Press, 2017.

Lynne M Stallings and Maryellen C MacDonald. It's not just the "Heavy NP": Relative phrase length modulates the production of heavy-NP shift. *Journal of psycholinguistic research*, 40(3):177–187, 2011.

Lynne M. Stallings, Maryellen C. MacDonald, and Padraig G O'Seaghdha. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language*, 39(3):392–417, 1998.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

Mihai Surdeanu and Christopher D. Manning. Ensemble models for dependency parsing: Cheap and good? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 649–652, Los Angeles, California, June 2010.

David Temperley. Dependency-length minimization in natural and artificial languages. *Journal of Quantitative Linguistics*, 15(3):256–282, 2008.

David Temperley. Minimization of dependency length in written English. *Cognition*, 105(2):300–333, 2007.

Lucien Tesnière. *Éléments de syntaxe structurale*. Klincksieck, 1959.

Lucien Tesnière. *Elements of structural syntax*. John Benjamins Publishing Company, 2015.

Juliette Thuilier. *Contraintes préférentielles et ordre des mots en français*. Ph.D. Thesis, Université Paris-Diderot - Paris VII, Sep 2012.

Juliette Thuilier, Gwendoline Fox, and Benoît Crabbé. Prédire la position de l'adjectif épithète en français : approche quantitative. *Lingvisticae Investigationes*, 35(1):28–75, 2012.

Harry Joel Tily. *The role of processing complexity in word order variation and change*. Ph.D. Thesis, Stanford University, 2010.

Ivan Titov and James Henderson. A Latent Variable Model for Generative Dependency Parsing. In Harry Bunt, Paola Merlo, and Joakim Nivre, editors, *Trends in Parsing Technology: Dependency Parsing, Domain Adaptation, and Deep Parsing*, pages 35–55. Springer Netherlands, Dordrecht, 2010.

Ivan Titov and James Henderson. A latent variable model for generative dependency parsing. In *Proceedings of the 10th International Conference on Parsing Technologies*,

IWPT '07, pages 144–155, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

Robert Truswell. Non-restrictive adjective interpretation and association with focus. In *Oxford Working Papers in Linguistics, Phonetics, and Philology*, volume 9, pages 133–154. 2005.

Hans Uszkoreit. A Framework for Processing Partially Free Word Order. In *Proceedings of the 21st Annual Meeting on Association for Computational Linguistics*, ACL '83, pages 106–112, Stroudsburg, PA, USA, 1983. Association for Computational Linguistics.

Marten Van Schijndel, Luan Nguyen, and William Schuler. An analysis of memory-based processing costs using incremental deep syntactic dependency parsing. *Proceedings of CMCL*, 2013.

Shravan Vasishth and Richard L. Lewis. Argument-Head Distance and Processing Complexity: Explaining both Locality and Antilocality Effects. *Language*, 82(4): 767–794, 2006.

Gabriella Vigliocco and Janet Nicol. Separating hierarchical relations and word order in language production: Is proximity concord syntactic or linear? *Cognition*, 68: B13–B29, 1998.

Dingquan Wang and Jason Eisner. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505, 2016.

Rui Wang and Yi Zhang. Sentence realization with unlexicalized tree linearization grammars. In *Proceedings of COLING 2012: Posters*, pages 1301–1310, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

Thomas Wasow. *Postverbal Behavior*. CSLI Publications, 2002.

Thomas Wasow. End-Weight from the Speaker's Perspective. *Journal of Psycholinguistic Research*, 26(3):347–361, May 1997.

Thomas Wasow and Jennifer Arnold. Post-verbal constituent ordering in english. In *Determinants of Grammatical Variation in English*, pages 119–154, 2003.

Linda Waugh. *A Semantic Analysis of Word Order*. EJ Brill, Leiden, 1977.

David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. Structured training for neural network transition-based parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 323–333, Beijing, China, July 2015. Association for Computational Linguistics.

Michael White and Rajakrishnan Rajkumar. Perceptron Reranking for CCG Realization. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, pages 410–419, 2009.

Michael White and Rajakrishnan Rajkumar. Minimal dependency length in realization ranking. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 244–255, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Daniel Wiechmann and Arne Lohmann. Domain minimization and beyond: Modeling prepositional phrase ordering. *Language Variation and Change*, 25(1):65–88, 2013.

Edwin Williams. Another argument that passive is transformational. *Linguistic Inquiry*, 13(1):160–163, 1982.

Bodo Winter. Linear models and linear mixed effects models in R with linguistic applications. *arXiv preprint arXiv:1308.5499*, 2013.

Marcin Woliński, Katarzyna Głowińska, and Marek Świdziński. A Preliminary Version of Skladnica—a Treebank of Polish. In Zygmunt Vetulani, editor, *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 299–303, Poznan, Poland, 2011.

Stephen Wu, Asaf Bachrach, Carlos Cardenas, and William Schuler. Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1189–1198. Association for Computational Linguistics, 2010.

Stefanie Wulff. A multifactorial corpus analysis of adjective order in English. *International Journal of Corpus Linguistics*, 8(2):245–282, 2003.

Hiroko Yamashita and Franklin Chang. "Long before short" preference in the production of a head-final language. *Cognition*, 81(2):B45–B55, 2001.

Victor H. Yngve. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 140:444–466, 1960.

Daniel Zeman, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. HamleDT: To Parse or Not to Parse? In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 23–25, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada, August 2017. Association for Computational Linguistics.