Structural and lexical factors in adjective placement in complex noun phrases across Romance languages

Kristina Gulordava
University of Geneva
Kristina.Gulordava@unige.ch

Paola Merlo
University of Geneva
Paola.Merlo@unige.ch

Abstract

One of the most common features across all known languages is their variability in word order. We show that differences in the prenominal and postnominal placement of adjectives in the noun phrase across five main Romance languages is not only subject to heaviness effects, as previously reported, but also to subtler structural interactions among dependencies that are better explained as effects of the principle of dependency length minimisation. These effects are almost purely structural and show lexical conditioning only in highly frequent collocations.

1 Introduction

One of the most widely observed characteristics of all languages is the variability in the linear order of their words, both across and within a single language. In this study, we concentrate on word order alternations where one structure can be linearised in two different ways. Consider, for example, the case when a phrasal verb (V + particle) has a direct object (NP), in English. Two alternative orders are possible: VP_1 = V NP Prt, and VP_2 = V Prt NP. If the NP is heavy, as defined in number of words or number of syllables, it will be frequently placed after the Prt, yielding the V-Prt-NP order. Compare, for instance *Call me up!* to *Call up the customer who called yesterday.* This tendency is also formulated as a Principle of End Weight, where phrases are presented in order of increasing weight (Wasow, 2002). Cases of heavy NP-shift (Stallings et al., 1998), dative alternation (Bresnan et al., 2007) and other alternation preferences among verbal dependents are traditionally evoked to argue in favour of the "heaviness" effect.

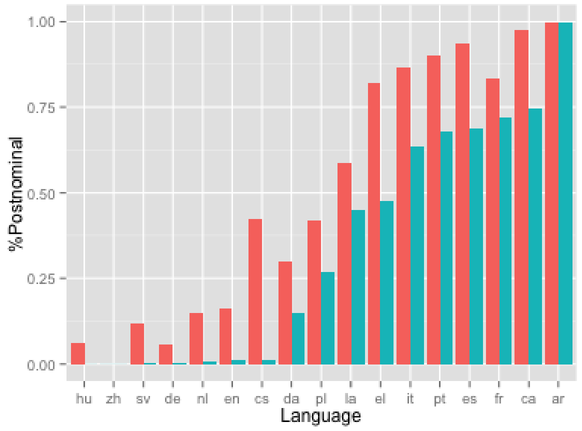


Figure 1: Percent of postnominal simple (green) and heavy (red) adjectives across seventeen languages.

In this work, we study the alternations in the noun-phrase domain, much less investigated in connection with the heaviness effect. Abeillé and Godard (2000) introduce the heaviness of adjective phrases as a principle explaining their postnominal placement compared to 'light' adjectives in French. Their observations have been recently confirmed in a corpus study by Thuilier (2012). Cross-linguistically, the data that we have collected across many languages and several families, presented in Figure 1, confirm the heaviness effect for adjectives[1]. By extracting relevant statistics from gold dependency annotated corpora, we can observe that heavy adjectives (adjective phrases of at least two words) appear more frequently postnominally than simple adjectives.

While the effect of size or heaviness is well-documented, this statistics is very coarse and it confounds various linguistic factors, such as types

[1] We use the following languages and treebanks: English, Czech, Spanish, Chinese, Catalan, German, Italian (Hajič et al., 2009), Danish, Dutch, Portuguese, Swedish (Buchholz and Marsi, 2006), Latin, Ancient Greek (Haug and Jøhndal, 2008), Hungarian (Csendes et al., 2005), Polish (Woliński et al., 2011), Arabic (Zeman et al., 2012), French (McDonald et al., 2013). The extraction is based on the conversion to the universal part-of-speech tags (Petrov et al., 2012).

Proceedings of the 19th Conference on Computational Language Learning, pages 247–257,
Beijing, China, July 30-31, 2015. ©2015 Association for Computational Linguistics

of adjectives, and annotation conventions of different corpora. From a typological perspective, the formulation needs to be refined from a preference of end weight to a preference for all elements being closer to the governing head: languages with Verb-Object dominant order tend to put constituents in 'short before long' order, while Object-Verb languages, like Japanese or Korean, do the reverse (Hawkins, 1994; Wasow, 2002). A more general explanation for the weight effect has been sought in a general tendency to minimise the length of the dependency between two related words, called Dependency Length Minimisation (DLM, Temperley (2007), Gildea and Temperley (2007)).

In this paper, we look at the structural factors, such as DLM, and lexical factors that play a role in adjective-noun word order alternations in Romance languages and the predictions they make on prenominal or postnominal placement of adjectives. We concentrate on a smaller set of languages than those shown in Figure 1 to be able to study finer-grained effects than what can be observed at a very large scale and across many different corpus annotation schemes. We choose Romance languages because they show a good amount of variation in the word order of the noun phrase.

The DLM principle can be stated as follows: if there exist possible alternative orderings of a phrase, the one with the shortest overall dependency length (DL) is preferred.

Consider, again, the case when a phrasal verb (verb + particle) has a direct object (NP). Two alternative orders are possible: VP$_1$ = V NP Prt, whose length is DL_1 and VP$_2$ = V Prt NP, whose length is DL_2. DL_1 is DL(V-NP)+DL(V-Prt) = |NP| + 1; DL_2 is DL(V-NP) + DL(V-Prt) = |Prt| + 1. If DL_1 is bigger than DL_2, then VP$_2$ is preferred over VP$_1$. Unlike the principle of End Weight, this explanation applies also to languages with a different word order than English.

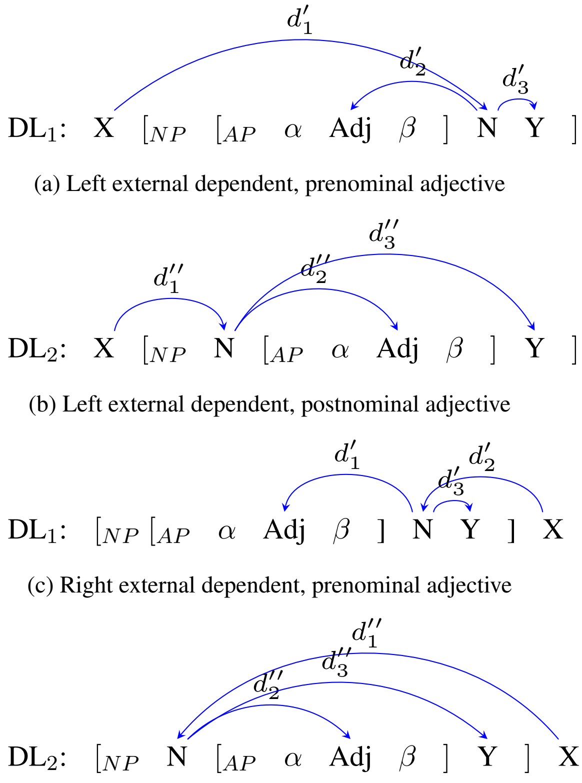
The observation that human languages appear to minimise the distance between related words is well documented in sentence processing (Gibson, 1998; Hawkins, 1994; Hawkins, 2004), in corpus properties of treebanks (Gildea and Temperley, 2007; Futrell et al., 2015), in diachronic language change (Tily, 2010). It is usually interpreted as a means to reduce memory load and support efficient communication. Dependency length minimisation has been demonstrated on a large scale in the verbal domain and at the sentence level, but has not yet been investigated in the more limited nominal domain, where dependencies are usually shorter and might create lighter processing loads that do not need to be minimised. In applying the general principle of DLM to the dependency structure of noun phrases, our goal is to test to what extent the DLM principle predicts the observed adjective-noun word order alternation patterns.

In this paper, we develop and discuss a more complex variant of a model described previously (Gulordava et al., 2015) and extend its analysis. First, we investigate whether the more complex DLM principle is necessary to explain our findings or if the simpler heaviness effect demonstrated for many languages in Figure 1 is sufficient. The answer is positive: the complexity introduced by DLM is necessary. Then, we develop a more detailed analysis of the only prediction of the model that is only weakly confirmed, showing that this result still holds under different definitions of dependency length. We also present an in-depth study to show that the DLM effect is structural, as assumed, and not lexical. While it is well-known that in French prenominal and post-nominal placement of adjectives is sometimes lexically-specific and meaning-dependent, this is not often the case in other languages like Italian, and does not explain the extent of the variation.

2 Dependency length minimisation in the noun phrase

In this section, we summarise the model in Gulordava et al. (2015). In the next section we propose a more complex model and study some factors in depth. Gulordava et al. (2015) consider a prototypical noun phrase with an adjective phrase as a modifier. They assume a simplified noun phrase with only one adjective modifier adjacent to the noun and two possible placements for an adjective phrase: post-nominal and prenominal. The adjective modifier can be a complex phrase with both left and right dependents (α and β, respectively). The noun phrase can have parents and right modifiers (X and Y, respectively). The structures for the possible cases are shown in Figure 2.

These structures correspond to examples like those shown in (1), in Italian (X='vede', Adj='bella', N='casa', Y= 'al mare').


d'_1
d'_2
d'_3
DL$_1$: X $[_{NP}$ $[_{AP}$ α Adj β] N Y]
d''_3
d''_1
d''_2
DL$_2$: X $[_{NP}$ N $[_{AP}$ α Adj β] Y]
d'_1
d'_2
d'_3
DL$_1$: $[_{NP}$ $[_{AP}$ α Adj β] N Y] X
d''_1
d''_3
d''_2
DL$_2$: $[_{NP}$ N $[_{AP}$ α Adj β] Y] X

(a) Left external dependent, prenominal adjective

(b) Left external dependent, postnominal adjective

(c) Right external dependent, prenominal adjective

(d) Right external dependent, postnominal adjective

Figure 2: Noun phrase structure variants and the dependencies relevant for the DLM calculation with right noun dependent Y.

	RightNP=Yes	RightNP=No
X=Left	$\lvert\beta\rvert - \lvert\alpha\rvert$	$2\lvert\beta\rvert + 1$
X=Right	$-3\lvert\alpha\rvert - 2$	$-2\lvert\alpha\rvert - 1$

Table 1: Dependency length difference for different types of noun phrases. By convention, we always calculate $DL_1 - DL_2$.

(1) a. ...vede la bella casa al mare.
('..sees the beautiful house at the sea')

b. ...vede la casa bella al mare.
('..sees the house beautiful at the sea')

c. La bella casa al mare è vuota.
('the beautiful house at the sea is empty.')

d. La casa bella al mare è vuota.
('the house beautiful at the sea is empty.')

The differences in dependency lengths predicted by DLM are summarized in Table 1. DLM makes predictions on adjective placement with respect to the noun —prenominal or postnominal— given the dependents of the adjectives, α and β, and given the dependent of the noun Y.

The column RightNP=No shows the dependency length difference for the two cases where the noun does not have a right dependent Y. Given that the calculation of DL differences is always calculated as $DL_1 - DL_2$, the fact that the cell (X=Left, RightNP=No) holds a positive value indicates that $DL_1 > DL_2$ and that the difference in length depends only on β and not on α. Conversely, the negative value of (X=Right, RightNP=No) shows that $DL_1 < DL_2$ and that the difference in length does not depend on β, but only on α. This is not intuitive: intuitively, one would expect that whether the Adjective is left or right of the Noun depends on the relative lengths of α and β, but instead if we look at all the dependencies that come into play for a noun phrase in a larger structure, the adjective position depends on only one of the two dependents. The table also shows that, on average, across all the cells, the weights of α are less than zero while the weights of β are greater than zero. This indicates that $DL_1 < DL_2$, which means that globally the prenominal adjective order is preferred.

DLM also makes predictions on adjective placement with respect to the noun given the dependents of the noun. Here the predictions of DLM are not intuitive. DLM predicts that when the external dependency is right (the dependency from the noun to its parent, X=right), then the adjective is prenominal, else it is postnominal. To spell this out, DLM predicts that, for example, we should find more prenominal adjectives in subject NPs than in NPs in object position. We discuss this odd prediction below.

Another prediction that will be investigated in detail is that when the noun has a right dependent, the prenominal adjective position is more preferred than when there is no right dependent, as evinced by the fact that the RightNP = Yes column is always greater than the RightNP = No column.

Gulordava et al. (2015) develop a mixed-effects model to test which of the fine-grained predictions derived from DLM are confirmed by the data provided by the dependency annotated corpora of five main Romance languages. The different elements in the DLM configuration are encoded as four factors: corresponding to the factors illustrated in Figure 2 and example (1), represented as binary or real-valued variables: *LeftAP* - the cumulative length (in words) of all left dependents of the adjective, indicated as α in Figure 2; *RightAP* - the cumulative length (in words) of all right dependents of the adjective, indicated as β in Figure 2; *ExtDep* - the direction of the arc from the noun to its parent X, an indicator variable; *RightNP* - the indicator variable representing the presence or absence of the right dependent of the noun, indicated as Y in Figure 2. [2]

Their findings partly confirm the predictions about adjective placement with respect to the noun given the adjective dependents. The DLM predictions about the position of the noun with respect to its parent are instead not confirmed. Finally, the prediction related to the presence of a right dependent of the noun on the placement of the adjective are confirmed.

In the next two sections, we replicate and investigate in more detail these results. First, we develop and discuss a more detailed model, where not only the factors, but also their interactions are taken into account. Then, we compare the predictions of the DLM model to the predictions of a simpler heaviness account, and confirm that the complexity of DLM is needed, as a simpler model based on heaviness of the adjective does not yield the same effects. Then, we discuss the external dependency factor, which, in the more complex model with interactions, is a significant factor. Finally, the RightNP factor is significant in the fitted model. The presence of a noun dependent on the right of the noun favours a prenominal placement, as predicted by DLM. We investigate the lexical aspects of this result in a more detailed case study.

3 Analysis of Dependency Minimisation Factors

The analysis that we develop here is based on the assumption that DLM is exhibited by the dependencies in the avalailable dependency-annotated corpora for the five Romance languages.

3.1 Materials: Dependency treebanks

The dependency annotated corpora of five Romance languages are used: Catalan, Spanish, Italian (Hajič et al., 2009), French (McDonald et al., 2013), and Portuguese (Buchholz and Marsi, 2006).

Noun phrases containing adjectives are extracted using part-of-speech information and dependency arcs from the gold annotation. Specifically, all treebanks are converted to coarse universal part-of-speech tags, using existing conventional mappings from the original tagset to the universal tagset (Petrov et al., 2012). All adjectives are identified using the universal PoS tag 'ADJ', whose dependency head is a noun, tagged using the universal PoS tag 'NOUN'. All elements of the dependency subtree, the noun phrase, rooted in this noun are collected. For all languages where this information is available, we extract lemmas of adjective and noun tokens. The only treebank without lemma annotation is French, for which we extract token forms.[3] A total of around 64'000 instances of adjectives in noun phrases is collected, ranging from 2'800 for Italian to 20'000 for Spanish.

3.2 Method: Mixed-Effects models

The interactions of several dependency factors are analysed using a logit mixed effect models (Bates et al., 2014). Mixed-effect logistic regression models (logit models) are a type of Generalized Linear Mixed Models with the logit link function and are designed for binomially distributed outcomes such as $Order$, in our case.

3.3 Factors and their interactions

While the original model in Gulordava et al. (2015) represents the four main factors involved in DLM in the noun phrase — α, β, RightNP and ExtDep — the predictions described above often mention interactions, which are not directly modelled in the original proposal. We introduce interactions, so that the model is more faithful to the DLM predictions, as shown in (2) and in Table 2. We do not take directly represent the interaction between the LeftAP and RightAP because in our corpora these two factors were both greater than zero in only 1$\%$ of the cases.

[2] In addition, to account for lexical variation, they include adjective tokens (or lemmas when available) as grouping variables introducing random effects. For example, the instances of adjective-noun order for a particular adjective will share the same weight value γ for the adjective variable, but across different adjectives this value can vary.

[3] During preprocessing, we exclude all adjectives and nouns with non-lower case and non-alphabetic symbols which can include common names. Compounds (in Spanish and Catalan treebanks), and English borrowings are also excluded. Neither do we include in our analysis noun phrases which have their elements separated by punctuation (for example, commas or parentheses) to ensure that the placement of the adjective is not affected by an unusual structure of the noun phrase.

Predictor	β	SE	Z	p
Intercept	-0.157	0.117	-1.33	0.182
LeftAP	2.129	0.183	11.63	$< .001$
RightAP	0.887	0.091	9.79	$< .001$
RightNP	-0.794	0.056	-14.24	$< .001$
ExtDep	-0.243	0.118	-2.06	0.039
RightNP:ExtDep	0.296	0.149	1.98	0.047
RightAP:RightNP:ExtDep	-0.639	0.353	-1.81	0.070

Random effects	Var
Adjective	1.989
Language	0.023

Table 2: Summary of the fixed and random effects in the mixed-effects logit model with interactions ($N = 15842$), shown in (2). Non-significant factors are not shown.

Model	Df	AIC	BIC	logLik	deviance	χ^2	Df	p
Without interactions	7	12137	12190	-6061.3	12123			
With interactions	14	12134	12241	-6052.9	12106	16.847	7	0.018*

Table 3: Comparison of the fits of two models: the model with interactions (2) and a simpler model without any interactions between the factors RightAP, LeftAP, RightNP and ExtDep.

(2) $$y_{ij} = (LeftAP + RightAP) \cdot RightNP \cdot \\ \cdot ExtDep \times \boldsymbol{\beta} + \gamma_{Adj_i} + \gamma_{Lang_j}$$

Contrary to the model without interactions (Gulordava et al., 2015), both the ExtDep factor and its interaction with the RightNP factor are significant. This interaction corresponds to the four different NP contexts presented in Table 1. Its significance, then, can be taken as preliminary confirming evidence for the distinction of these contexts, as predicted by DLM. A direct comparison of the two models, with and without interactions, shows, however, that the effects of these interactions are rather small (Table 3). The log-likelihood test shows that the model with interactions fits the data significantly better ($\chi^2 = 16.8, p = 0.02$), but the comparison of the Bayesian Information Criterion scores of the two models — criterion which strongly penalises the number of parameters — suggests that the model without interactions should be preferred.

3.4 Comparison of DLM and heaviness model

Dependency length minimisation was introduced, as mentioned in the introduction, to better explain processing effects at the sentence level for which heaviness accounts were inadequate. However, noun phrases are small and relatively simple domains. We ask, then, if a model is sufficient where the AP is not divided into LeftAP and RightAP, but holistically represented by the size of the whole AP.

Specifically, a simpler Heaviness model does not make a difference between left and right dependent of adjectives: all heavy adjectives are predicted to move post-nominally. Heaviness would also not predict the interaction between placement and the existence of a noun dependent to the right.

We compare, then, two minimally different models. Since neither the external dependency factor nor the interactions were shown to be highly significant, we compare a simplified DLM model shown in (3) to a model where AP is represented only by its heaviness (number of words) as in (4).

(3) $$y_{ij} = LeftAP \cdot \beta_{LAP} + RightAP \cdot \beta_{RAP} \\ + RightNP \cdot \beta_{RNP} + \gamma_{Adj_i} + \gamma_{Lang_j}$$

(4) $$y_{ij} = SizeAP \cdot \beta_{HV} + RightNP \cdot \beta_{RNP} \\ + \gamma_{Adj_i} + \gamma_{Lang_j}$$

The DLM model that distinguishes LeftAP from RightAP in (3) fits the data better than a model where AP is represented only by its heaviness as in (4), as can be seen in Table 4 and from the difference in AIC values of two models ($\Delta AIC = 146$). This result confirms that the complexity introduced by DLM minimisation is needed, and confirms DLM as a property of language, also in noun phrases. The main conceptual difference between heaviness accounts and DLM

	Df	AIC	BIC	logLik	deviance	χ^2	Df	p
Model with *SizeAP*	5	12518	12557	-6254.1	12508			
Model with *LeftAP*, *RightAP*	6	12372	12418	-6179.8	12360	148.5	1	$< .001$

Table 4: Comparison of the simplified DLM model in (3) and the heaviness model in (4).

accounts resides in the fact that the former do not take into account the structure and the nature of the context of the heavy element, while DLM does. This model comparison shows that adjective placement is not independent of its context.

Prediction for External Dependencies The expected effect of the external dependency of the noun predicted by the DLM is borne out only marginally. This factor predicts a difference between noun phrases that precede their head, for example subjects, and noun phrases that follow their head, for example objects. The prediction is unexpected, while the result that the factor is not highly significant less so, as it is not immediately clear why nouns preceding heads should behave differently from nouns that follow heads.

A possible explanation for this mismatch of the predictions and the observed data patterns lies in the assumptions underlying the DLM principle. We have assumed a definition of dependency length as the number of words between the head and the dependent, as found in the corpus annotation. Our data are annotated using a content-head rule, which assumes that the noun is the head of the noun phrase. Hawkins (1994), in his well-developed variant of DLM, postulates that minimisation occurs on the dependencies between the head and the *edge* of the dependent phrase. For noun phrases, the relevant dependencies will span between the determiner which unambiguously defines the left edge of the noun phrase and the head of NP (e.g., a verb). The predictions of Hawkins' theory for adjective placement will therefore differ from the DLM predictions based on our definition. As can be observed from Figure 2, the d'_1 and d''_1 dependencies to the left edge of the NP will be of equal length in cases (a) and (b) (similarly to d'_2 and d''_2 in cases (c) and (d)). The external dependency is predicted therefore not to affect the resulting adjective placement, as observed in the data. This result lends weak support to a theory where in this case the relevant dependency is between the parent and the edge of the dependent.

A question remains of what dependencies are minimised when the noun phrase does not have a determiner and the left edge of the noun phrase is ambiguous.[4] This issue is difficult to test in practice in our corpora. First, there are many more cases (84% versus 16%) with left ExtDep (X is on the right, e.g. for object NPs) than with right ExtDep (X is on the left, e.g. for subjects) in Romance languages. This is because all of them, except French, can optionally omit subjects. Moreover, the function of the NP, subject or object, and therefore the ExtDep variable, correlates with the definiteness of the NP. NPs in object position take an article 75% of time while NPs in subject position take an article 96% of time. Consequently, NPs without articles and on the left of the head are observed only 135 times in our data sample (across all languages). This small number of cases did not allow us to develop a model.

4 In-depth study of the RightNP dependency factor

The most novel result of the model in Gulordava et al. (2015), extended here to the more complex model (2) concerns the interaction between the adjective position and the RightNP. This effect would not be predicted by a heaviness explanation and even in the DLM framework it is surprising that minimisation should apply to such a short dependency. We investigate this interaction in more detail and ask two questions: is this effect evenly spread across different nouns and adjectives or is it driven by some lexical outliers? what are the lexical effects of the noun and its dependent? We analyse a large amount of data constructed to be a representative sample of adjective variation for several nouns (around thirty for each language) and very many adjectives and investigate noun phrases with a right dependent introduced by the preposition 'de/di'[5].

[4] In one of his analyses, Hawkins claims that adjectives define unambiguously the left edge of the NP, but this assumption is controversial.

[5] For Italian, the preposition is 'di', while for other three languages it is 'de'. We do not consider complex prepositions such as 'du' in French or 'do' in Portuguese.

4.1 Data extracted from large PoS-tagged corpora

We extract the data by querying automatically a collection of corpora brought together by the SketchEngine project (Kilgarriff et al., 2014). This web-interface-based collection allows partially restricted access to billion-word corpora of Italian (4 billions of words), French (11 billions), Spanish (8.7 billions) and Portuguese (4.5 billions). The corpora are collected by web-crawling and automatically PoS-tagged. A similar Catalan corpus was not available through this service.

First, we define the set of seed nouns that will be queried. For each language, we use our treebanks to find the list of the two-hundred most frequent nouns which take the ‘di/de’ preposition as a complement. A noun has ‘di/de’ as its right dependent if there is a direct head-dependent link between these elements in the gold annotation. Nouns in the list which could be ambiguous between different types of parts of speech are replaced manually. We randomly sample around thirty nouns, based on the percentage of their co-occurrence with ‘di/de’. Given the list of seed nouns, we automatically queried the four corpora with simple regular patterns containing these nouns to extract cases of prenominal and postnominal noun-adjective co-occurrences.[6]

For each noun, we collected a maximum of 100’000 matches for each of the two patterns, which is the SketchEngine service limit. These matches include left and right contexts of the pattern and allow to extract the token following the pattern, which can be ‘di/de’ or nothing.

We modeled the data using the Logit mixed effect models, with the $Order$ as a response variable, one fixed effect (Di) and nouns and adjectives as random effects. We fit the maximal model with both slope and intercept parameters, as shown in model (5).

$$\begin{aligned}(5)\qquad y &= Di \cdot (\beta_{Di} + \beta_{Adj_i} + \beta_{Noun_j})\\ &+ \gamma_{Adj_i} + \gamma_{Noun_j}\end{aligned}$$

We fit our models on a sample of data of around 200’000 instances of adjective-noun alternations for each language, equally balanced for noun phrases with $Di = True$ and $Di = False$.

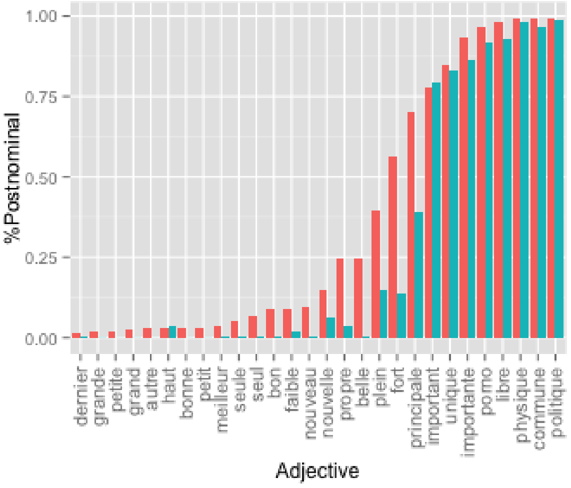


Figure 3: Percent postnominal placement for the thirty most frequent adjectives in French. (Noun phrase has a right ‘de’-complement (green) and it does not (red).

4.2 Results

The data shows that the Di effect is small, but highly significant for all languages. The resulting values are similar: for French $\beta_{Di} = -0.84$, Portuguese $\beta_{Di} = -0.95$, Italian $\beta_{Di} = -1.14$ and Spanish $\beta_{Di} = -1.65$ (all $p < 0.001$).

Figure 3 illustrates the Di effect for French (cumulative for all nouns). We observe that most of the adjectives appear more frequently prenominally in noun phrases with a ‘de’ complement than in noun phrases without a ‘de’ complement (green columns are smaller than corresponding red columns). Importantly, we observe a very similar picture cross-linguistically for all four languages and for the adjective alternation across the majority of the nouns (if considered independently), as shown in Figure 4.

This result agrees with our predictions, and shows that DLM effects show up even in short spans, where they are not necessarily expected. If a postnominal adjective intervenes between the noun and the dependent, the dependency length increases only by one word (with respect to the noun phrase with the prenominal adjective). Our results nevertheless suggest that even such short dependencies are consistently minimised. This effect is confirmed in all languages.

4.3 Lexical effects on adjective placement

One of the lexical factors that could play a confounding role for the prenominal placement of ad-

[6] Our patterns were of the type ‘[tag=”ADJ”] *noun*’ and ‘*noun* [tag=”ADJ”]’, where the tag field is specified for the PoS tag of a token. In our case, ‘A.’ was the tag for adjectives in , and ‘ADJ’ in Italian, French and Spanish.

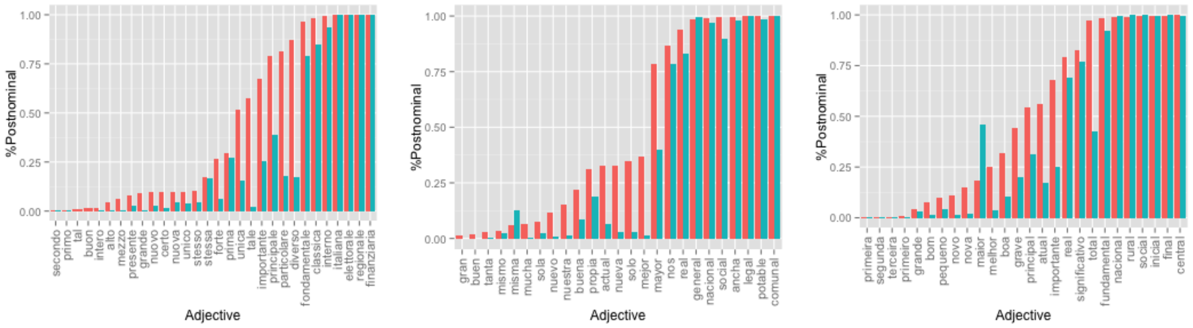


Figure 4: Percent postnominal placement for the thirty most frequent adjectives in Italian, Spanish, and Portuguese (in this order). (Noun phrase has a right 'de/di'-complement (green) and it does not (red).

jectives in Di constructions is the strength of the 'Noun + di/de + Complement' collocation. For example, in the French compound 'taux de chomage' ('unemployement rate') the placement of an adjective after the compound — 'taux de chomage important' — is preferred in our corpora to the adjacent postnominal placement ('taux important de chomage'). In our analysis, we do not extract these types of post-NP adjectives. From this perspective, a drop in the percentage of postnominal adjectives in 'di' cases could indicate that adjectives prefer not to intervene between nouns and their complements. We hypothesize that this dependency is more strongly minimised than other dependencies in the noun phrase because of this strong lexical link.

We confirm that the *Di* effect is an interaction of the DLM principle and lexical properties of compounds by a further preliminary analysis of collocations. From the French data, we selected a subset with the most frequent 'Noun + de + Complement' sequences (10 for each seed noun) and a subset with infrequent sequences (100 random decomplements for each seed noun). We assume that the frequency of the sequence is an indicator of the collocational strength, so that highly frequent sequences are collocations while low frequency sequences are non-collocational combinations. The first subset has a proportion of 78% prenominal and 22% postnominal adjectives, while the second subset has 61% prenominal and 39% postnominal adjectives. We confirm, then, that in the frequent collocations there is a substantial lexical effect in adjective placement. However, we also observe a preference of prenominal placement for the infrequent 'Noun + de + Complement' sequences that are not collocational combinations, since prenominal placement is still much higher than what is observed for French adjectives, on average (46% prenominal and 54% postnominal in our sample of data). These numbers suggest that the Di effect reported in the previous section is not a result of mere lexical collocation effects and that, for low frequency combinations at least, DLM is at play.

A different kind of lexical effect is shown in Figure 5. Here we plot the percent postnominal placement of the adjective, if the noun has a complement introduced by *di (of), che (that), per (for),* in Italian. We notice that adjective placement is no longer as majoritarily prenominal for the right dependent introduced by *che* and *per* as it is for *di*. The main difference between *di* (of) and *che* (that), *per* (for) is that the former introduces a PP that is inside the NP that selects it, while *che* and *per* usually do not, they are adjuncts, or infinitivals or clauses. In the linguistic literature, this is a distinction between arguments and adjuncts of the noun and it is represented structurally. This distinction is, then, a lexically-induced structural distinction, and not simply a collocation.

5 Related work

Our work occupies the middle ground between detailed linguistic investigations of weight effect in chosen constructions of well-studied languages and large scale demonstrations of the dependency length minimisation principle.

Gildea and Temperley (2007) demonstrated that DLM applies for the dependency annotated corpora in English and German. They calculate random and optimal dependency lengths for each sentence given its unordered dependency tree and compare these values to actual dependency lengths. English lengths are shown to be close to optimal, but for German this tendency is not as clear. A recent study of Futrell et al. (2015) applies

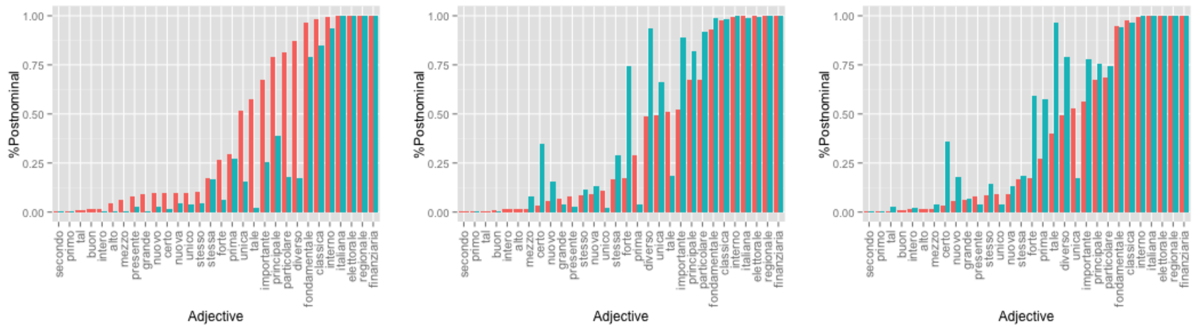


Figure 5: Percent postnominal placement for thirty most frequent adjectives in Italian, followed by function word *di, che, per*, in this order. (Noun phrase has a right 'di/per/che'-complement (green) and it does not (red)).

this analysis on a large-scale, for more than thirty languages that have dependency treebanks. Their results also confirm the correspondence between the dependency annotation and the experimental data, something that has been reported previously (Merlo, 1994; Roland and Jurafsky, 2002).

Much work in theoretical linguistics addresses the adjective-noun order in Romance languages. Such work typically concentrates on lexico-semantic aspects of adjective placement (Cinque, 2010; Alexiadou, 2001). In our work, we account for the strong lexical prenominal or postnominal preferences of adjectives by including them as random effects in our models.

Closest to our paper is the theoretical work of Abeillé and Godard (2000) on the placement of adjective phrases in French and recent corpus-based work by Fox and Thuilier (2012) and Thuilier (2012). Fox and Thuilier (2012) use a dependency annotated corpus of French to extract cases of adjective-noun variation and their syntactic contexts. They model the placement of an adjective as a lexical, syntactic and semantic multi-factorial variation. They find, for example, that phonologically heavy simple adjectives tend to be postnominal. This result highlights the distinction between phonological weight and syntactic weight, a topic which we do not address in the current work.

6 Conclusion

In this paper, we have shown that differences in the prenominal and postnominal placement of adjectives in the noun phrase across five main Romance languages is not only subject to heaviness effects, but to subtler dependency length minimisation effects. These effects are almost purely structural and show lexical conditioning only in highly frequent collocations.

The subtle interactions found in this work raise questions about the exact definition of what dependencies are minimised and to what extent a given dependency annotation captures these distinctions. Future work will investigate more refined definitions of dependency length minimisation, that distinguish different kinds of dependencies with different weights.

Acknowledgements

We gratefully acknowledge the partial funding of this work by the Swiss National Science Foundation, under grant 144362. Part of this work was conducted during the visit of The first author to the Labex EFL in Paris and Alpage INRIA group. We thank Benoit Crabbé for the fruitful discussions during this period.

References

Anne Abeillé and Daniele Godard. 2000. 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.

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

Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker, 2014. *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. 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.

Sabine Buchholz and Erwin Marsi. 2006. 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. Association for Computational Linguistics.

Guglielmo Cinque. 2010. *The Syntax of Adjectives: A Comparative Study*. MIT Press.

Dóra Csendes, János Csirik, Tibor Gyimóthy, and András Kocsor. 2005. The Szeged treebank. In *Text, Speech and Dialogue*, pages 123–131. Springer.

Gwendoline Fox and Juliette Thuilier. 2012. 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.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-Scale Evidence of Dependency Length Minimization in 37 Languages. (Submitted to Proceedings of the National Academy of Sciences of the United States of America).

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Daniel Gildea and David Temperley. 2007. 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.

Kristina Gulordava, Paola Merlo, and Benoit Crabbé. 2015. 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: Short Papers (ACL'15)*.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, pages 1–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dag T. T. Haug and Marius L. Jøhndal. 2008. 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.

John A Hawkins. 1994. *A performance theory of order and constituency*. Cambridge University Press, Cambridge.

John A. Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford linguistics. Oxford University Press, Oxford, UK.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97. Association for Computational Linguistics.

Paola Merlo. 1994. A corpus-based analysis of verb continuation frequencies for syntactic processing. *Journal of Psycholinguistic Research*, 23(6):435–457.

Slav Petrov, Dipanjan Das, and Ryan T. McDonald. 2012. 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.

Douglas Roland and Daniel Jurafsky. 2002. Verb sense and verb subcategorization probabilities. In Paola Merlo and Suzanne Stevenson, editors, *The lexical basis of sentence processing: Formal, computational, and experimental issues*. John Benjamins.

Lynne M Stallings, Maryellen C MacDonald, and Padraig G O'Seaghdha. 1998. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language*, 39(3):392–417.

David Temperley. 2007. Minimization of dependency length in written English. *Cognition*, 105(2):300–333.

Juliette Thuilier. 2012. *Contraintes préférentielles et ordre des mots en français*. Ph.D. Thesis, Université Paris-Diderot - Paris VII, Sep.

Harry Joel Tily. 2010. *The role of processing complexity in word order variation and change*. Ph.D. Thesis, Stanford University.

Thomas Wasow. 2002. *Postverbal Behavior*. CSLI Publications.

Marcin Woliński, Katarzyna Głowińska, and Marek Świdziński. 2011. 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.

Daniel Zeman, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. 2012. 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. European Language Resources Association (ELRA).