Beauty Before Age?
Applying Subjectivity to Automatic English Adjective Ordering

Felix Hill
Dept. of Theoretical & Applied Linguistics
and Computer Laboratory
University of Cambridge
Cambridge CB3 9DA, UK
fh295@cam.ac.uk

Abstract

The preferred order of pre-nominal adjectives in English is determined primarily by semantics. Nevertheless, Adjective Ordering (AO) systems do not generally exploit semantic features. This paper describes a system that orders adjectives with significantly above-chance accuracy (73.0%) solely on the basis of semantic features pertaining to the cognitive-semantic dimension of *subjectivity*. The results indicate that combining such semantic approaches with current methods could result in more accurate and robust AO systems.

1 Introduction

As a significant body of linguistic research has observed (see e.g. Quirk et al. (1985)), English pre-nominal adjective strings exhibit subtle order restrictions. Although example (2), below, does not represent a clear-cut violation of established grammatical principles, it would sound distinctly unnatural to native speakers in the majority of contexts, in contrast to the entirely unproblematic (1).

(1) He poked it with a long metal fork
(2) ? He poked it with a metal long fork

The problem of determining the principles that govern Adjective Ordering (henceforth, AO) in English has been studied from a range of academic perspectives, including philosophy, linguistics, psychology and neuroscience. AO is also of interest in the field of Natural Language Processing (NLP), since a method that consistently selects felicitous orders would serve to improve the output of language modeling and generation systems.

Previous NLP approaches to AO infer the ordering of adjective combinations from instances of the same, or superficially similar, combinations in training corpora (Shaw & Hatzivassiloglou, 1999) (Malouf, 2000), or from distributional tendencies of the adjectives in multiple-modifier strings (Mitchell, 2009) (Dunlop, Mitchell, & Roark, 2010). Such methods are susceptible to data sparseness, since the combinations from which they learn are rare in everyday language.

By contrast, the approach taken here determines AO based on semantic features of adjectives, guided by the theoretical observation that the cognitive notion of *subjectivity* governs ordering in the general case (Adamson, 2000). The semantic features developed are each highly significant predictors of AO, and they combine to classify combinations with 73.0% accuracy. These preliminary results indicate that semantic AO systems can perform comparably to existing systems, and that classifiers exploiting semantic and direct evidence might surpass the current best-performing systems.

2 Previous research

The subtle nature of human ordering preferences makes AO a particularly challenging NLP task. In perhaps the first specific attempt to address the problem, Shaw and Hatzivassiloglou (1999) apply a *direct evidence* method. For a given adjective combination in the test data, their system searches a training corpus and selects the most frequent ordering of that combination. Because there is no basis to determine the order of adjective combinations that are not in the training data, Shaw and Hatzivassiloglou extend the domain of the classifi-

Proceedings of the NAACL HLT 2012 Student Research Workshop, pages 11–16,
Montréal, Canada, June 3-8, 2012.

er by assuming transitivity in the order relation, increasing the coverage with only a small reduction in accuracy. Nevertheless, the system remains highly dependent on the domain and quantity of training data. For example, accuracy is 92% when training and test data are both within the medical domain but only 54% in cross-domain contexts.

Malouf (2000) combines a direct evidence approach with an alternative method for extending the domain of his classifier. His system infers the order of unseen combinations from 'similar' seen combinations, where similarity is defined purely in terms of morphological form. The method works by exploiting a degree of correlation between form and order (e.g. capital letters indicate nominal modifiers, which typically occur to the right).

Mitchell (2009) applies a less 'direct' approach, clustering adjectives based on their position in multiple-modifier strings. Although Mitchell's classifier requires no direct evidence, data sparseness is still an issue because the strings from which the system learns are relatively infrequent in everyday language. Dunlop et al. (2010) apply Multiple Sequence Alignment (MSA), a statistical technique for automatic sequence ordering, which, as with Malouf's system, quantifies word-similarity based solely on morphological features. Despite the greater sophistication of these more recent approaches, Mitchell et al. (2011) showed that a simple n-gram (direct evidence) classifier trained on 170 million words of New York Times and Wall Street Journal text and tested on the Brown Corpus (82.3% accuracy) outperforms both the clustering (69.0%) and MSA (81.8%) methods.

Wulff (2003) uses Linear Discriminant Analysis (LDA) to quantify the effects of various potential AO correlates, and confirms that semantic features are better predictors than morphological and syntactic features. The features, extracted from the 10-million word Spoken British National Corpus (BNC) and weighted by LDA, combine to predict unseen adjective orders with 72% accuracy.

Wulff's study is unique in applying semantics to the problem, although her focus is theoretical and several features are implemented manually. The next section describes the theoretical basis for a fully-automated semantic approach to AO that could help to resolve the issues of data sparsity and domain dependence associated with the direct evidence methods described above.

2.1 The subjectivity hypothesis

Although phonetic, morphological and syntactic factors influence AO in specific contexts, there is consensus in the theoretical literature that semantics is the determining factor in the general case (see Quirk et al. (1985) for further discussion). Several semantic theories of AO make use of the cognitive linguistic notion of *subjectivity* (Quirk et al. 1985; Hetzron, 1978; Adamson 2000). Subjectivity in this context refers to the degree to which an utterance can or cannot be interpreted independently of the speaker's perspective (Langacker, 1991). For example, the deictic utterance (3) is more subjective than (4) since its truth depends on the speaker's location at the time of utterance.

(3) James is sitting across the table
(4) James is sitting opposite Sam

In relation to AO, Quirk et al, Hetzron and Adamson each support some form of the *subjectivity hypothesis:* that more subjective modifiers generally occur to the left of less subjective modifiers in prenominal strings. For example, in (5) the adjective *big* tells us about the relation between the car and the speaker's idea of typical car size. This ascription is less objectively verifiable than that of car color, so *big* occurs further from the head noun. The position of *oncoming* in (6) reflects the high inherent subjectivity of deictic modifiers.

(5) A big red Italian car (BNC)
(6) An oncoming small black car (BNC)

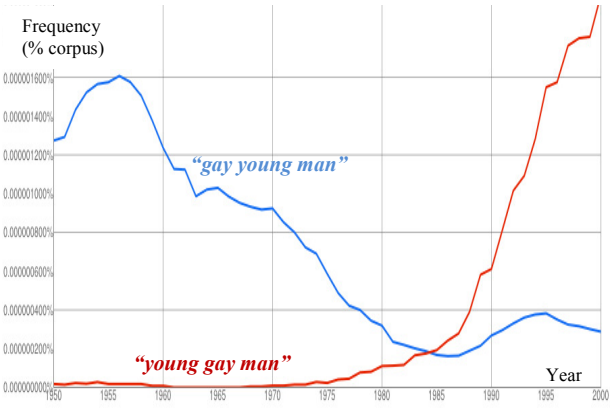

Figure 1: Diachronic variation of preferred AO

To illustrate a process of changing AO preferences that can be explained in a compelling way by the subjectivity hypothesis, the 1 trillion-word Google n-Gram Viewer was queried (Figure 1). The two

lines indicate the frequency of the strings *'gay young man'* and *'young gay man'* in the Corpus from 1950 to 2000, as the pre-eminent meaning of *gay* evolved from the subjective *merry* to the categorical, well-defined *homosexual*. As the graph shows, this reduction in subjectivity has been accompanied by a marked increase in the tendency of *gay* to appear closer to the noun in such strings.

3 System design

The AO system described below applies the theoretical findings presented above by extracting from training data various subjectivity features of adjectives and applying this information to classify input orderings as correct or incorrect.[1] System operation and evaluation consisted of 5 stages.

***Extracting feature profiles*:** The 200 highest-frequency adjectives in the BNC were extracted. Following Wulff (2003, p. 6), three items, *other*, *only* and *very* were removed from this list because they occur in right-branching structures. For the remaining adjectives, a 'profile' of feature values (c.f. Table 1, below), was extracted from 24 million words (*Sections A-C*) of the written BNC.

***Generating gold-standard orderings*:** From the 197 adjectives, 19,306 unordered pairs $\{A_1, A_2\}$ were generated. The bigram frequencies of the strings $[A_1, A_2]$ and $[A_2, A_1]$ were then extracted from the 1 billion-word Google n-gram Corpus. From this data, the 12,000 pairs $\{A_1, A_2\}$ with the largest proportional difference in frequency between $[A_1, A_2]$ and $[A_2, A_1]$ were selected.

***Defining test and training sets*:** A set of 12,000 ordered triples $[A_1, A_2, \delta_{[A_1, A_2]}]$ was generated, where $\delta_{[A_1, A_2]}$ is an indicator function taking the value 1 if $[A_1, A_2]$ is the preferred ordering in the Google corpus and 0 if $[A_2, A_1]$ is preferred. Some of the triples were re-ordered at random to leave an equal number of preferred and dispreferred orderings in the data. These triples were populated with feature profiles, to create vectors

$$[f_1^{A_1}, \ldots, f_n^{A_1}, f_1^{A_2}, \ldots f_n^{A_2}, \delta_{[A_1, A_2]}]$$

where $f_k^{A_i}$ is the value of the k^{th} feature of the adjective A_i, and n is the total number of features. The set of vectors was then randomly partitioned in the ratio 80:20 for training and testing respectively.

***Training the classifier*:** A logistic regression was applied to the set of training vectors, in which the first *2n* elements of the vectors were independent variables and the final element was the dependent variable. Logistic regression has been shown to be preferable to alternatives such as Ordinary Least Squares and LDA for binary outcome classification if, as in this case, the independent variables are not normally distributed (Press & Wilson, 1978).

***Evaluation*:** Performance was determined by the number of pairs in the test data correctly ordered by the classifier. Steps 3-5 were repeated 4 times (5-fold cross-validation), with the scores averaged.

3.1 The Features

Of the features included in the model, COMPARABILITY and POLARITY are shown to correlate with human subjectivity judgments by Wiebe and colleagues (see e.g. Hatzivassiloglou & Wiebe, 2000). The remainder are motivated by observations in the theoretical literature.

MODIFIABILITY: Gradable adjectives, such as *hot* or *happy*, tend to be more subjective than prototypically categorical adjectives, such as *square* or *black* (Hetzron, 1978). Unlike categorical adjectives they admit modification by intensifiers (Paradis, 1997). Therefore, the feature MODIFIABILITY is defined as the conditional probability that an adjective occurs immediately following an intensifier given that it occurs at all.[2]

$$Modifiability(A) = \frac{\sum_{m \in M} freq([m, A])}{freq(A)}$$

$M = \{degree\ modifiers\}$
$[x, y]$ *is the bigram 'word x followed by word y'*

COMPARABILITY: Gradable adjectives also have comparative and superlative forms, whereas prototypically categorical adjectives do not. Given the association between gradability and subjectivity, the feature COMPARABILITY is defined as the probability of an adjective occurring in comparative or superlative form given it occurs at all.

[1] The system operates on adjectival and nominal modifiers but not on articles, determiners, degree modifiers and other non-adjectival pre-modifiers.

[2] The set of intensifiers is taken from (Paradis, 1997).

$$Comparability(A) = \frac{freq(\bar{A}) + freq(\bar{\bar{A}})}{freq(A) + freq(\bar{A}) + freq(\bar{\bar{A}})}$$

$\bar{A} = comparative\ form\ of\ A \quad \bar{\bar{A}} = superlative\ form\ of\ A$

PREDICATIVITY: Adjectives can be applied in both attributive (‘*the red car’*), and predicative (‘*the car is red’*) constructions. Bolinger (1967) suggests that predicative constructions are conceptualized more dynamically or temporarily than attributive constructions. Since dynamic properties are generally ascribed more subjectively than permanent properties (Langacker, 1991), Bolinger’s intuition implies an association between subjectivity and predicative constructions. Indeed, many objective modifiers sit uncomfortably in predicative contexts, as shown by (7) and (8).

(7) I live in a brick house
(8) ? The house I live in is brick

The feature PREDICATIVITY is therefore defined as the probability that an adjective occurs in a predicative construction given that it occurs at all. The measure is implemented by counting the number of times the adjective immediately follows some form of an English copula verb.[3]

$$Predicativiy(A) = \frac{\sum_{c \in C} freq([c, A])}{freq(A)}$$

$C = set\ of\ English\ copula\ verbs\ in\ all\ inflected\ forms$

POLARITY: An adjective is said to be *polar* if it typically attributes a positive (*kind, healthy, strong*) or negative (*poor, selfish, rotten*) characteristic. Semi-supervised methods for automatically detecting adjective polarity have been developed (Hatzivassiloglou & McKeown, 1997), and applied to subjectivity analysis by Wiebe (2000). POLARITY is implemented as a binary feature, whose value depends on whether or not the adjective appears in a list of 1,300 polar adjectives extracted by Hatzivassiloglou & Mackeown.

$$Polarity(A) = \begin{cases} 1 & if\ A \in \boldsymbol{P} \cup \boldsymbol{N} \\ 0 & if\ A \notin \boldsymbol{P} \cup \boldsymbol{N} \end{cases}$$

$\boldsymbol{P} = \{adjectives\ labelled\ as\ positive\}$
$\boldsymbol{N} = \{adjectives\ labelled\ as\ negative\}$

[3] The copula verbs list was compiled manually by the author.

ADVERBIABILITY: Quirk (1985, p 1339) notes that evaluative adjectives tend to develop derived adverbial forms, whereas more objective adjectives do not. For example, *nice*, *beautiful* and, *careful* correspond to the adverbs *nicely*, *beautifully*, and *carefully*, whereas no such derived forms exist for the more objective adjectives *male*, *English* and *brown*. The ADVERBIABILITY of an adjective is defined as the ratio of derived adverbial forms to total base and adverbial forms in the corpus.

$$Adverbiability(A) = \frac{freq(A^*)}{freq(A) + freq(A^*)}$$

$A^* = adverbial\ form\ derived\ from\ A$

NOMINALITY: Wullf (2003) reports statistical evidence that more ‘noun-like’ modifiers appear closer to the head in modifying strings. Combinations such as ‘*bread knife’* or ‘*police car’*, often analyzed as noun-noun compounds rather than modifier/noun combinations, represent the clearest such examples. Amongst more prototypical adjectives, some, such as *green*, or *male* have nominal senses (‘*village green’*, ‘*unidentified male’*), whereas others do not. Separately, Hatzivassiloglou and Wiebe (2000) report a statistical correlation between the number of adjectives in a text and human judgments of subjectivity. These observations suggest that adjectives are inherently more subjective than nouns, and further that noun-like ‘behavior’ might indicate relative objectivity within the class of adjectives. Consequently, the feature NOMINALITY is defined, following Wulff, as the probability that an adjective is tagged as a noun given that it is tagged as either an adjective or a noun. It is the only feature that is expected to exhibit an inverse correlation with subjectivity.

$$Nominality(A) = \frac{freq(A_n)}{freq(A_a) + freq(A_n)}$$

$A_n = adjective\ A\ tagged\ as\ noun$
$A_a = adjective\ A\ tagged\ as\ adjective$

	new	***good***	***old***	***different***	***local***
MODIF	0.0010	0.0529	0.0208	0.0887	0.0004
COM	0.0079	0.4881	0.2805	0.0011	0.0045
PRED	0.0100	0.1018	0.0289	0.0806	0.0069
POL	0.0000	1.0000	0.0000	0.0000	0.0000
ADV	0.0220	0.0008	0.0000	0.0318	0.0478
NOM	0.2900	0.0999	0.0113	0.0000	0.0212

Table 1: Example feature profiles

4 Results

The performance of the classifier is promising with respect to the intuition that semantic features can be usefully applied to AO systems. A chi-square test reveals the features collectively to be highly significant predictors of AO ($\chi^2 = 2257.25,\ p < 0.001^{***}$). Once trained, the system orders unseen combinations in the test data with accuracy of 73.0%, as detailed in Table 2. This figure is not directly comparable with previous work because of differences in the evaluation framework.

		Predicted		
		Training Data		
		Incorrect	Correct	% Correct
Observed	Incorrect	2773	1637	62.9
	Correct	1120	4101	78.5
	Overall%			71.4
		Test Data		
		Incorrect	Correct	% Correct
	Incorrect	696	370	65.3
	Correct	270	1031	79.2
	Overall%			73.0

Table 2: Overall results of model cross-validation

It is notable that the accuracy of the classifier rises to 86.2% when the test data is hand-picked as the 3000 pairs for which the strength of ordering preference is highest.[4] This suggests that the approach could be particularly effective at detecting highly unnatural combinations. Moreover, the performance when tested on the 3000 (unseen) pairs with the lowest ordering preference is 70.1%, indicating the potential to cope well with marginal cases and rare combinations.

As Table 3 shows, all features apart from COMPARABILITY are statistically significant predictors in the model ($p < 0.001^{***}$). In addition, the mean value of each feature over adjectives in first position A_1 differs significantly from the mean over adjectives in second position A_2 ($t \geq 28.07$ in each case, $df = 11{,}283$). Whilst relatively the weakest predictor, COMPARABILITY in isolation does predict AO at above-chance cy ($58.7\%,\ p < 0.001^{***}$)

. Its low significance in the overall model reflects its high level of interaction with other features; in particular, MODIFIABILITY (Pearson Correlation: .367, $p < 0.001^{***}$). The relative magnitude of the model coefficients is not informative, since the measurement scale is not common to all features. Nevertheless, the negative regression coefficient of NOMINALITY confirms that this feature correlates inversely with distance from the noun.

Feature	Regression Coefficient	Predictor Significance	Performance in Isolation	Comparison of A_1 / A_2 Means
MODIF	5.205	.000	62.9%	0.000
COM	.177	.381	58.7%	0.000
PRED	3.630	.000	68.6%	0.000
POL	.339	.000	60.4%	0.000
ADV	1.503	.000	62.8%	0.000
NOM	-.405	.000	58.4%	0.000

Table 3: Influence of individual features

To test the influence of the training corpus size on system performance, features were extracted from BNC *Section A* (7 million words) rather than *Sections A-C* (24 million words) in a separate experiment. This adjustment resulted in a reduction in classifier accuracy from 73.0% to 71.4%, indicating that performance could be significantly improved by training on the full BNC or even larger corpora. Further improvements could be achieved through the combination of semantic and 'direct' features. To illustrate this, the feature LEFTTENDENCY, a measure of the likelihood that an adjective occurs immediately to the left of another adjective in the training data, was added. This adjustment raised the classifier accuracy from 73.0% to 76.3%. It should also be noted that many of the features in the current system are extracted via measures that approximate syntactic dependency with bigram context. It is an empirical question whether the additional complexity associated with more precise measures (for example, applying dependency parsing) would be justified by performance improvements.

[4] The 3000 pairs for which the proportional preference for one ordering over another in the Google n-Gram corpus is highest and for which the total frequency of the pair exceeds 500.

5 Conclusion

This paper has tested the efficacy of applying automatic subjectivity quantification to the problem of AO. The reported results highlight the utility of such semantically oriented approaches. Although direct comparison with existing systems was beyond the scope of this study, exploratory analyses suggested that a refined version of this system might compare favorably with reported benchmarks, if trained on a corpus of comparable size.

Nevertheless, the comparatively weak performance of the present system on previously seen examples ('underfitting', see Table 2) is strong evidence that six features alone are insufficient to capture the complexity of ordering patterns. Therefore, beyond the adjustments discussed above, the next stage in this research will evaluate the effects of combining semantic features with direct evidence in a single system. Other future work might apply subjectivity features to cluster adjectives into classes pertinent to AO, perhaps in combination with independent distributional measures of semantic similarity. Finally, the approach presented here for English AO could have applications across languages, and may also be applicable to related tasks, such as ordering binomials[5], parsing noun phrases ('*wild animal hunt'* vs. '*wild birthday party'*) and selecting thematically appropriate modifiers for a given head noun.

Some interesting theoretical insights also emerge as a corollary to the results of this study. The supposition that gradability, polarity, adverbiability, predicativity and 'nouniness' can be associated, either positively or negatively, with subjectivity, was confirmed. Moreover, the performance of the classifier lends support to the status of subjectivity as a determining principle of AO, and an important dimension of adjective semantics in general. As such, the reason we say *beautiful English rose*, (c.240,000 direct matches on Google) and not *English beautiful rose* (c.2,730) is because beauty is in the eye of the beholder, whereas nationality, evidently, is not.

[5] Binomials are noun or adjective combinations separated by coordinating conjunctions, such as *tired and emotional* and *salt and pepper.* Quirk et al. (1985, p. 1342) observe connections between binomial ordering and AO.

Acknowledgments

Thanks to Anna Korhonen, Paula Buttery and Sylvia Adamson for helpful guidance and comments.

References

Adamson, S. 2000. Word Order Options and Category Shift in the Premodifying String. In O. Fischer, *Pathways of Change: Grammaticalization in English* (pp. 39-66). Amsterdam: John Benjamins.

Bolinger, D. 1967. Adjectives in English: Attribution and Predication. *Lingua 18* , 1-34.

Dunlop, A., Mitchell, M. & Roark, B. 2010. Prenominal Modifier Ordering via Multiple Sequence Alignment. *2010 Annual Conference of the North American Chapter of the ACL (HLT-NAACL 2010)*.

Hatzivassiloglou, V. & McKeown, K. 1997. Predicting the Semantic Orientation of Adjectives. *Annual Meeting Assoc. Comp.Ling. ACL '97,* 174-181.

Hatzivassiloglou, V. & Wiebe, J. 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. *International Conference on Computational Linguistics, COLING- '00.*

Hetzron, R. 1978. On the Relative Order of Adjectives. In I. H. (Ed.), *Language Universals*. Tübingen: Narr.

Langacker, R. 1991. *Foundations of Cognitive Grammar.* Stanford, CA: Stanford University Press.

Malouf, R. 2000. The Order of Prenominal Adjectives in Natural Language Generation. *Proc. 38th Annual Meeting, Assoc. Comp. Linguistics, ACL '00*, 85–92.

Mitchell, M. 2009. Class-based Ordering of Prenominal Modifiers. *Proc.12th European Workshop, Nat.Lang. Generation, ENLG '09*, 50–57.

Mitchell, M. Dunlop, A. & Roark, B. 2011. Semi-Supervised Modeling for Prenominal Modifier Ordering. *Proc. 49th Annual Meeting of the Assoc. Comp. Ling., ACL '11,* 236–241.

Paradis, C. 1997. *Degree Modifiers of Adjectives in Spoken British English.* Lund: Lund University Press.

Press, S. J. & Wilson, S. 1978. Choosing Between Logistic Regression and Discriminant Analysis. *Journal of American Statistical Association*, 699-705.

Quirk, R. Greenbaum, A. Leech, G. & Svartvik, J. 1985. *A Comprehensive Grammar of the English Language.* London: Longmans.

Shaw, J. & Hatzivassiloglou, V. 1999. Ordering Among Premodifiers. *Proc. 37th Annual Meeting, Association of Computational Linguistics, ACL '99* , 135–143.

Wiebe, J. 2000. Learning Subjective Adjectives from Corpora. *Proc. 17th National Conference on Artificial Intelligence (AAAI-2000)*.

Wulff, S. 2003. A Multifactorial Analysis of Adjective Order in English. *International Journal of Corpus Linguistics*, 245–282.