Originally published in: Flanders, Julia/Jannidis, Fotis (Eds.):
The shape of data in digital humanities. Modeling texts and text-based resources. -
London [et al.]: Routledge, 2019. Pp. 236-249. (Digital Research in the Arts and Humanities)

Linguistic and computational modeling in language science

Elke Teich
Peter Fankhauser

Linguistics is concerned with modeling language from the cognitive, social, and historical perspectives. When practiced as a science, linguistics is characterized by the tension between the two methodological dispositions of rationalism and empiricism. At any point in time in the history of linguistics, one is more dominant than the other. In the last two decades, we have been experiencing a new wave of empiricism in linguistic fields as diverse as psycholinguistics (e.g., Chater et al., 2015), language typology (e.g., Piantidosi and Gibson, 2014), language change (e.g., Bybee, 2010) and language variation (e.g., Bresnan and Ford, 2010). Consequently, the practices of modeling are being renegotiated in different linguistic communities, readdressing some fundamental methodological questions such as: How to cast a research question into an appropriate study design? How to obtain evidence (data) for a hypothesis (e.g., experiment vs. corpus)? How to process the data? How to evaluate a hypothesis in the light of the data obtained? This new empiricism is characterized by an interest in *language use in context* accompanied by a commitment to computational modeling, which is probably most developed in psycholinguistics, giving rise to the field of "computational psycholinguistics" (cf. Crocker, 2010), but recently getting stronger also in corpus linguistics.

The predominant domain of corpus linguistics is language variation, aiming at statements on relative differences/similarities between linguistic varieties (time periods, registers, genres). Corpus analysis is thus comparative by nature; technically, this involves comparing probability distributions of (sets of) linguistic features (e.g., the relative frequency of passive vs. active voice in narrative vs. expository genres) and assessing whether they are significantly different or not. Here, descriptive statistical techniques come into play but also language modeling and machine learning methods (e.g., clustering, latent semantic analysis, or Bayesian modeling). Similarly, corpus processing—that is, preparing text material for analysis—relies on computational models, for example, for annotation. What is important to note here is that processing and analysis are broken up into different steps, each using a different computational *micro-model* that takes care of a specific task (e.g., labeling linguistic units in annotation) and consists of a *descriptive component* (set of allowed labels) and an *analytic* or *algorithmic component* (procedure by which labels are assigned).

In this chapter, we focus on one such task—part-of-speech tagging (in linguistic terms: grammatical word classification)—and the class of computational models addressing this task. In so doing, we discuss the differences between models constructed by human observation and computational models induced from corpus data. The major points we would like to stress here are that all models (human or machine-made) (a) are approximations and will never achieve the "perfect" description, and (b) start from a set of prior assumptions about modeling. Regarding computational models, it is then up to the human user to decide whether the model assumptions are reasonable and whether the degree of descriptive accuracy achieved is good enough for a given purpose of analysis.

The remainder of the chapter is organized as follows. We briefly introduce the basic workflow adopted in corpus-based research and relate its components to the relevant types of data as well as to the kinds of theoretical sources that inform the micro-models in different stages of processing/analysis. We then introduce a standard linguistic model for parts-of-speech, including a historical perspective, and discuss in more detail the role of modeling assumptions in computational approaches to part-of-speech tagging. In the concluding section we discuss implications of the perspectives on modeling presented in this article for modeling in the language- and text-oriented humanities more widely.

1 Types of data and theoretical sources for modeling

We assume the now common technical conception of a corpus-linguistic workflow as a processing pipeline distinguishing between processing of raw and primary data and analysis of primary data for obtaining secondary data (cf. Himmelmann, 2012, from the perspective of language documentation). Raw data can be recordings of spoken language (audio/video) or written text documents. Primary data can be transcriptions of audio/video or plain text or annotated text with structural mark-up, tokenization, part-of-speech tagging, and so forth. Secondary data can be descriptive statements—for example, dictionary entries or grammatical descriptions, but also frequency distributions and their interpretations. While raw data are unique, primary and secondary data are not: there are always alternative ways of processing raw data and different kinds of primary as well as secondary data can be derived from it. Here, the difference between primary and secondary data is sometimes not clear-cut. However, primary data is typically closer to the linguistic signal than secondary data, and secondary data requires primary data as input (e.g., in order to calculate a probability distribution of the word classes in a text or corpus, it needs to be tagged first in terms of parts-of-speech).

Each data type (raw, primary, secondary) is associated with a particular processing stage and requires specific methods for processing. Procedures to get from raw data to primary data may involve full text digitization, text normalization, sentence segmentation, tokenization, lemmatization, morphological analysis, part-of-speech tagging, and syntactic parsing, but also manual annotation (e.g., annotation of semantic roles and relations). Together, these processing steps enable the derivation of primary data from raw data. Each of them has its own

underlying micro-model in the sense defined above, that is, a descriptive and an algorithmic model for a specific processing task. Depending on the nature of the task, these models are theoretically informed by linguistic theory, probability theory and/or information theory. The steps in deriving secondary data (e.g., a probability distribution) again follow particular micro-models that define input and output, based on formal grammar (e.g., regular expressions for corpus query) and descriptive statistics or data mining (for assessing probability distributions). Figure 1 summarizes the commonly adopted processing steps in relation to data types and theoretical sources for modeling.

Importantly, the performance of each micro-model can be tested separately by measuring how well it fits a given data set and predicts the behavior of new data. Again, it is important to note that 100 percent accuracy will never be attained, but knowing about model quality, we can decide how the error rate may affect the next steps in processing or analysis.

Why would it be interesting to compare linguistic modeling as carried out by humans and computational modeling as carried out by machines? There is necessarily a gap between a model that is designed for computation and a model that is designed for human consumption. While both require determining the object of modeling, making explicit the descriptive categories to be used and providing criteria for assigning categories, the goals of modeling may be different, the task itself may need to be differently defined, and consequently the models themselves (both descriptive and analytic) will be different. A model that was

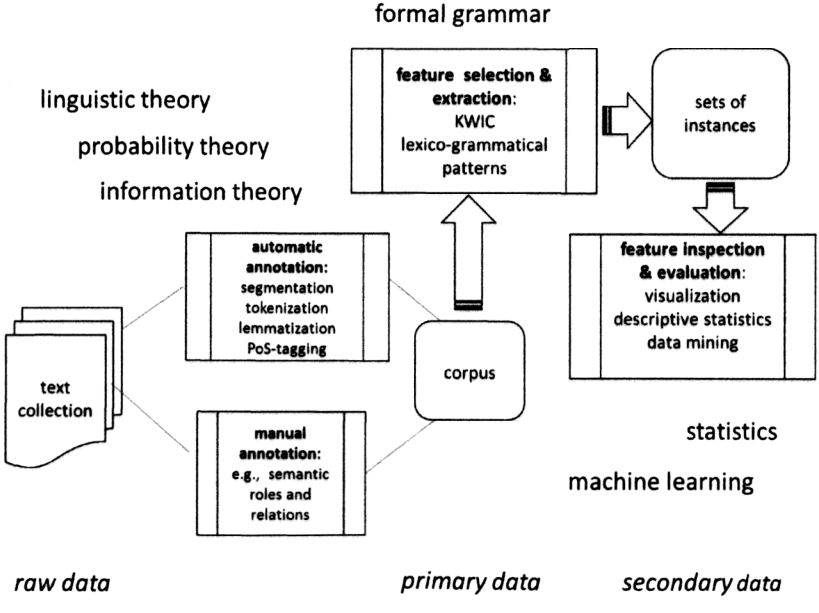

Figure 1 Processing pipeline, data types, and theoretical sources for modeling.

designed for/by humans can thus typically not be straightforwardly applied by a machine, and vice versa models designed for/by machines are not necessarily easy to interpret for humans.

For illustration, we look at the classification of words into parts-of-speech (PoS)—in computational terms, part-of-speech tagging. We discuss the relation of a linguistic model of parts-of-speech as we would typically find it in a standard grammar (e.g., Quirk et al., 1985 for English) and computational models for automatically assigning PoS tags to strings.

2 Grammatical classification of words

Traditionally, linguistics is concerned with classification—that is, abstracting from observations of linguistic instances to classes. The goal of classification is to come up with a descriptive model of a given object (syllable, word, clause, etc.).

A standard work for many centuries after it appeared, the *Téchnē Grammatikē*, a description of Ancient Greek, is attributed to Dionysius Thrax (*c.* 100BC). The two basic *units of description* identified were the *sentence* and the *word*. The word was defined as the smallest meaning-bearing unit that was not further decomposable (there was awareness of morphology, but no term for it yet). Observing the behavior of the word, Thrax came up with the following eight word classes (parts-of-speech): noun, verb, participle, article, pronoun, preposition, adverb and conjunction (cf. Robins, 1997, p. 41). The properties on the basis of which these classes were distinguished were mainly to do with the internal properties of words—for example, nouns (onoma) and verbs (rhema) were distinguished on the basis of case inflection (+ / –) as the sole distinctive feature. Interesting cases are the recognition of the participle as a separate class and the non-recognition of the adjective as a separate class. In Ancient Greek, the participle is both case- and tense-inflected—so it has properties of nouns and verbs; the adjective is very similar to nouns in morphology and syntax, so adjectives and nouns were subsumed under one class. Each class is described in terms of its attributes referring to grammatically relevant differences in the forms of words, essentially what we refer to today as grammatical categories (such as gender, case, tense, voice, mood, person etc.) and some syntactic criteria (e.g., preposition placed before other words, adverb modifying a verb).

Thrax's classification can be called a micro-model in that it focuses on one particular constituent of language, the word, so Thrax must have been aware of the necessity to break down a complex object (language) into manageable subparts. Furthermore, Thrax's classification is an instance of a descriptive model; in today's computational linguistic terminology, it is a part-of-speech tag set. We cannot be entirely sure about how Thrax arrived at this model, but he definitely proceeded in an empirical fashion. The data he used was taken from written texts by accepted authors of the time and he must have inspected this data very closely. Regarding the analytic part of his model, we do not have much evidence. Generally, Thrax will have applied Aristotelian methods of classification,

but there is no explicit account of the criteria he used. As mentioned above, the criteria he will have applied are to a large degree to do with the internal properties of words and to a lesser extent syntactic and distributional.

A contemporary classification of words, which basically applies across languages, assumes eight word classes—noun, verb, adjective, pronoun, article, adverb, interjection, conjunction, and preposition—and is only slightly different from Thrax's classification (interjection has been added, participle has been removed). The differences are due to more insights into the properties of words, notably morphology, but also their syntactic behavior and distributional properties. The latter are crucial in defining criteria for grammatical disambiguation. Consider examples (1) and (2).

(1) The file has been deleted.
(2) Can you file the report?

This kind of ambiguity is particularly common in English because noun-to-verb conversion is very productive. In isolation, the word *file* is ambiguous between a noun (1) and a verb (2), but in the context of a preceding article (1) *file* can only be a noun and in the context of a preceding pronoun, it can only be a verb (2). We will come back to the importance of syntactic context in the following sections on computational modeling.

While there are clearly many languages in the world that have only been partially described (or not described at all), methodologically grammatical word classification counts as a solved task in modern linguistics: the principles of word classification and the procedures linguists use to detect word classes are course-book knowledge. They include substitution tests, syntactic tests (e.g., reordering of elements) as well as distributional information. One general insight from the experiences in linguistic modeling is that any model will be approximate: linguistic classes are typically gradient, some members being at the core of a class exhibiting all defining features, others carry only some of the defining features and are at the periphery of a class.

In summary, traditionally the goal of modeling in linguistics is to come up with a descriptive model of a linguistic object. For modeling purposes, language is broken up into manageable parts that are linguistically relevant (such as words). The task of modeling consists of detecting the classes (descriptive model) and providing criteria for distinguishing between them (analytic model).

3 Part-of-speech tagging

Part-of-speech (PoS) tagging belongs to the most commonly applied types of corpus processing. Words are a linguistically relevant unit for grammatical and semantic study; but even if we are not specifically interested in studying words, PoS are a very useful abstraction from strings that we can use—for example, in corpus search. The importance of PoS tagging was recognized quite early on in corpus-based research in the late 1960s and considerable efforts went into

```
Emma/NNP Woodhouse/NNP ,/, handsome/JJ ,/, clever/JJ ,/,
and/CC rich/JJ ,/, with/IN a/DT comfortable/JJ home/NN
and/CC happy/JJ disposition/NN ,/, seemed/VB to/TO unite/VB
some/DT of/IN the/DT best/JJS blessings/NNS of/IN
existence/NN ;/;

word classes: NNP= proper noun, JJ = adjective, CC =
coordinating conjunction, IN = preposition, DT =
determiner, NN = noun singular, VB = verb, TO = to, JJS =
adjective superlative, NNS = noun plural
```

Figure 2 A sample of Jane Austen's *Emma*, tagged with parts of speech.

manual tagging—for example, the work on the Brown and LOB corpora carried out by Francis and Kučera (1982). Nowadays, automatic PoS tagging achieves a very high accuracy (95–97 percent), so that its output can serve as input for further processing and analysis.

Consider a sample sentence (the first sentence from Jane Austen's *Emma*) tagged with parts-of-speech with the Stanford tagger and associated tag set (Toutanova and Manning, 2000):

(3) Emma Woodhouse, handsome, clever, and rich, with a comfortable home and happy disposition, seemed to unite some of the best blessings of existence;

We can see here a number of differences to traditional linguistic word classification. First, not only words are being tagged, but also punctuation marks. Second, the tag set is more extensive than the commonly used set of word classes, partly because it encodes some grammatical categories (e.g., number in nouns) as well as semantic categories (e.g., proper vs. common nouns).

In contrast to traditional linguistic word classification, the overall goal of part-of-speech tagging is to assign tags to *all tokens* in a text rather than just to words; also, if it is easy to cover grammatical categories in tagging, they are encoded in tags (in linguistic classification, grammatical classes and categories are strictly kept separate). Therefore, the descriptive models of word classification and PoS tagging differ.

The analytic part of a PoS tagger can either be rule-based (e.g., Brill, 1992) or statistically based (e.g., Schmid, 1994; Toutanova and Manning, 2000). In statistical approaches, modeling is based on conditional probabilities and can follow a supervised or an unsupervised approach.

3.1 Supervised part-of-speech tagging

Conditional probabilities are widely used in statistical *language modeling* (Rosenfeld, 2002). Based on word n-grams they calculate the probability of an upcoming word *w* based on the context of the previous words (typically one, two, or three words)—that is,

$$P(w_n \mid w_{n-i} \ldots w_{n-1})$$

In contrast to plain word n-gram based approaches, part-of-speech tagging includes an abstraction step in that it calculates the probability $P(w_n \mid c_n)$ of a word w_n, given its class c_n, which in turn is conditioned on the sequence of preceding classes: $P(c_n \mid c_{n-i} \ldots c_{n-1})$. The task of part-of-speech tagging is thus to predict the class of a given word based on the sequence of preceding classes, and possibly also its following classes, which can be derived from the above probabilities on the basis of Bayes's rule:[1]

$$P(c_n \mid w_n, \ldots) \propto P(w_n \mid c_n) * P(c_n \mid c_{n-i} \ldots c_{n-1}) \qquad \text{(Equation 1)}$$

The underlying modeling assumption is that a given linguistic event (e.g., a word) in a sequence of linguistic events is dependent on previous (and following) events in the sequence. More specifically, words are assumed to be generated by the following two-stage stochastic process:

For every word w_n

(1) choose a class c_n from the class-sequence distribution

$$P(c_n \mid c_{n-i} \ldots c_{n-1})$$

(2) choose the word from the class-word distribution

$$P(w_n \mid c_n)$$

In supervised part-of-speech tagging, these probabilities can be directly estimated from a training corpus. Of course, this simplified model does not really capture all aspects of what constitutes a word class. In the simplest approach, $P(w_n \mid c_n)$ just memorizes the frequencies of observed words in a class, and does not take into account, for example, morphological characteristics of words, which presumably served as a basis for Thrax's classification. Moreover, $P(c_n \mid c_{n-i} \ldots c_{n-1})$ is only characterized in terms of rather coarse syntactic categories, and only considers a context of fixed length, thereby disregarding long-range syntactic dependencies. In fact, state-of-art part-of-speech taggers use more elaborate features that do take into account morphology and other characteristics of words in context.

Still, this type of model constitutes the conceptual foundation for algorithmic part-of-speech tagging (see Hidden Markov Models (HMMs) as described, for example, in Manning and Schütze (1999: Chapters 9 and 10)). The original idea goes back to Markov (1913) who applied it to the initial 20,000 characters of Pushkin's *Evgeni Onegin* in order to predict vowel-consonant sequences—a task very similar to part-of-speech tagging.

3.2 Unsupervised part-of-speech tagging

Hidden Markov models can also be deployed in an unsupervised fashion. Rather than estimating the class-word distributions $P(w_n \mid c_n)$ and the class sequence n-grams $P(c_n \mid c_{n-i} \ldots c_{n-1})$ from labeled training examples, Equation 1 is used to estimate latent classes. Thus, the modeling assumption is fully reduced to the generative stochastic process described above, there exists no descriptive model in form of a given set of word classes; only the number of classes and the length of context (order of the HMM) is given.

One of the first approaches to this end was introduced by Brown et al. (1992). In the analysis below, we use the Bayesian approach introduced by Goldwater and Griffiths (2007), which uses annealed Gibbs sampling based on Equation 1 to approximate the class-word and class-sequence distributions. It is instructive to compare this to the Gibbs sampling equation used for topic modeling (Steyvers and Griffiths, 2007, cf. Underwood, this volume):

$$P(z \mid w, \ldots) \propto P(w \mid z) * P(z \mid d) \qquad \text{(Equation 2)}$$

where $P(w/z)$ is the topic-word distribution, and $P(z/d)$ is the document-topic distribution. The underlying generative process is very similar:

For every word w

(1) choose a topic z from the document-topic distribution $P(z/d)\}$
(2) choose the word from the topic-word distribution $P(w/z)$.

Thus, the essential difference between topic models and unsupervised Hidden Markov models is that the document-topic distribution considers the bag of topics of an entire document as context, whereas the class-sequence distribution $P(c_n \mid c_{n-i} \ldots c_{n-1})$ considers only the local, ordered class context. Griffiths et al. (2004) describe an approach that combines these two latent modeling approaches.

For illustration, we have applied this approach to the Brown/LOB family of corpora (Francis and Kučera, 1982), comprising about 4.7 million tokens of British and American English. Table 1 lists the PoS classes using a HMM of order 2 (two preceding classes) and assuming 25 classes. All classes except *sentence marker* are latent classes, their grouping, description, and labeling with tags from the Penn tagset (Marcus et al., 1993) are derived by qualitative analysis of the class-word distributions and most frequent class sequences. The column labeled "Ent." gives the entropy of the class-word distributions measured in bits, with low values indicating a closed class consisting of only few different words, and high values indicating an open class. The resulting class-word distributions illustrate the strengths but also the limitations of the underlying modeling assumptions: Some major syntactic categories (*nouns*, *verbs*, *adjectives*, *prepositions*, *conjunctions*, *personal pronouns*) are identified quite well. *Nouns* are differentiated into two noun singular classes, noun plural, proper nouns (titles), and countables. *Personal pronouns* are roughly differentiated by their grammatical

Table 1 Latent PoS-Classes with a HMM of order 2

%	*Ent.*	*Tags*	*Description*	*Top 5 Words*				
4.9	0.6	SENT	sentence marker	.	?	:	!	“
5.3	0.9	PUNC	punctuation	,	;	“	than*	‘
7.2	11.3	NN	noun singular (1)	time	man	way	year	day
3.3	10.8	NN	noun singular (2)	part	number	sense	end	kind
5.1	9.5	NNS	noun plural	people	men	women	children	things
2.2	11.6	NP	title, first name	mr	john	mrs	sir	miss
2.3	9.5	NN/CD	countable	years	2	1	per	3
4.6	5.4	PP	personal pron. (1)	he	i	it	she	they
2.5	10.5	PP	personal pron. (2)	it	him	them	me	her
2.0	10.9	JJ*	adjective (1)	good	important	true	possible	better
3.3	10.3	JJ*	adjective (2)	first	new	own	same	old
10.3	2.7	DT/JJ/PP$	before noun	the	a	his	an	this
4.4	10.7	DT/JJ/PP$	bef. Noun plur.	These	their	other	two	new
8.0	4.4	IN	preposition (1)	in	to	for	with	on
3.1	0.8	IN	preposition (2)	of	between	that*	is*	called*
3.9	8.0	VV(D)	verb lexical	is	was	said	are	had
2.6	9.0	VB/VD/VV	verb infinitive	be	do	make	see	get
3.7	10.4	V*G/V*N	verb participle	made	used	going	taken	given
4.4	4.3	V*/MD	verb auxiliary	‘*	is	was	had	are
1.6	1.2	TO/MD	before verb inf.	To	will	would	can	may
3.4	5.8	RB/VB/VH	before verb	s	be	not	been	have
2.9	7.5	RP/PP	after verb	t	it	out	up	him
2.9	4.8	CC	conjunction (1)	“*	but	if	then	however
3.3	2.0	CC	conjunction (2)	and	or	but	as	so
2.7	5.8	W*/CC	rel. clause/conj.	That	which	as	if	when

case, *verbs* by four major grammatical categories, and *conjunctions* by their position in sentence: (1) in sentence initial position, and (2) within a sentence. Punctuations are also clearly recognized as an individual class.

Some other classes are less well separated: *Determiners* get mixed up with other classes that may occur before a noun, such as demonstrative articles and adjectives. *Auxiliary verbs* contain almost 25 percent of apostrophes arising from either possessive "s" or verb contractions ("don't")—these get separated into individual classes, when working with a larger number of classes. Most prominently, *adverbs* are not separated well at all, but get mixed up with other classes that may occur before or after a verb. Here, neglecting morphological characteristics in estimating the class-word distributions seems to strike particularly hard. Regular adverbs in English are signaled by the suffix "ly," which may be due to the difficulty of recognizing adverbs by their syntactic context alone.

The class *before verb infinitive* is particularly interesting: 87 percent of its probability mass is accounted for by "to"+infinitive, and the next most frequent words are modal verbs typically followed by an infinitive. However, as the examples in Table 2 show, "to"+infinitive (right) is well distinguished from "to" as a preposition (left). The latent syntactic context—prepositions cannot be followed by an infinitive—serves well to disambiguate between the two uses of "to."

In a similar vein, Table 3 gives examples of the disambiguation between "be/have/do" as lexical verbs (left) and as auxiliary verbs. Also in these examples, the right context of the verb (determiner vs. verb or verb participle) enables the disambiguation. The two examples "it is possible" and "it is difficult" constitute interesting borderline cases. Strictly speaking "is" followed by an adjective is to be classified as a lexical verb, but apparently the model cannot pick up the subtle difference between "is" followed by a participle (e.g., "she is educated") vs. followed by an adjective. Note that supervised part-of-speech taggers using the Penn tagset classify all occurrences of "be/have/do" as "VB/VH/VD"—that is, the distinction between lexical and auxiliary verbs is not regarded at all.

In summary, unsupervised approaches to part-of-speech tagging are a natural generalization of supervised approaches. This does not only hold for part-of-speech tagging; also other generative models, such as topic models, can span the continuum between fully supervised and fully unsupervised modes of operation

Table 2 Disambiguation of "to" into preposition vs. "to"+infinitive

Occ.	*RP/PP*	*PP*	*DT*	*Occ.*	*VV(D)*	*TO*	*VB/VD/VV*
232	back	to	the	94	seems	to	be
201	up	to	the	90	seemed	to	be
149	on	to	the	58	was	to	be
142	away	from	the	57	had	to	do
133	up	in	the	53	appears	to	be

Table 3 Disambiguation of lexical vs. auxiliary verbs

Occ.	*PP/EX*	*VV(D)*	*DT*	*Occ.*	*PP*	*VB/VD/VH/MD*	*V*G/V*N/VV*
602	there	was	a	67	it	is	possible*
585	there	is	a	62	he	was	going
575	it	was	a	61	we	have	seen
422	it	is	a	59	i	would	like
352	it	was	the	58	it	is	difficult*

(Ramage et al., 2011). However, because part-of-speech tagging, at least for well-resourced languages such as English, is so well understood, it allows to understand the limitations of overly simplistic generative models by means of qualitative and quantitative analysis, as exemplified above by the apparent importance of the morphological regularity of adverbs in English for their classification. When applied to less-resourced languages or more specialized language varieties, the weak modeling assumptions can also serve for discovering specific patterns and classes that cannot be captured with a fixed class vocabulary.

4 Conclusion and envoi

We have discussed selected aspects of modeling language from the linguistic and the computational perspectives. Ultimately, linguistics is interested in generalizations about language as a cognitive and a social system. Modeling linguistic data–that is, describing it and generalizing from it—is a means to this end. Analysis is broken down into manageable subparts, for which descriptive and analytic micro-models are devised that are increasingly computationally supported. We currently experience a push towards empirical approaches, even in areas that have hitherto been committed to a rationalist perspective (cf. section 12.1). With this development comes the need to reflect more on analytic processes and to model these processes. Providing such models on the part of computational linguistics/computer science and getting accustomed to working with them on the part of linguistics paves the ground for a new linguistic empiricism that is computationally informed.

What can be learned from these experiences for the humanities more widely? At a first glance, the computationally informed empirical turn in linguistics seems to widen the gap between it and the humanities. However, this would be adopting a rather traditionalistic perspective (cf. Bod, 2013, for related discussions). In fact, with the recent developments in digital humanities, it may even turn out to be the contrary. First, linguistic data are humanistic data—that is, they are contextualized in time and space. (Computational) linguists realize more and more that extra-linguistic context is extremely important in the analysis of language and the interpretation of linguistic acts (cf. Halevy et al., 2009). But how exactly context can be modeled beyond the immediate context of words remains an open question (cf. also Church, 2011, for a discussion in computational linguistics).

In the digital humanities, in turn, recent advances in computational processing show that there may be more information in the linguistic signal than humanists may have hitherto assumed. This is shown in studies of text using machine learning in fields as diverse as history (e.g., Hinrichs et al., 2014, uncovering historical facts in nineteenth-century global economy) and literary theory (see e.g., Chambers and Jurafsky, 2009 or Mimno et al., 2014, for approaches automatically detecting narrative structure). What characterizes this direction of research is, first, the awareness of the need to stick closely to the linguistic signal in order to capture patterns that can subsequently be interpreted at more abstract levels and, second, the readiness to apply methods and techniques from machine learning to detect patterns that would otherwise remain hidden.

What should also be kept in mind is that computational language models are linguistically informed—that is, they take (some of) their model assumptions from linguistics. Grammatical word classification is a very well-understood task in linguistics and computational models are thus fairly well linguistically informed. Discussing a concrete model of unsupervised part-of-speech tagging using distributional criteria alone, we have seen that fairly good analysis results may be achieved on this basis, but if additional linguistic criteria were taken into consideration (e.g., morphology), such a model would clearly perform better. Moreover, while part-of-speech tagging is a very well-understood task, there are many other kinds of phenomena that are much less well understood, such as the encoding of writer/speaker attitude in text, principles of text structuring, or genre classification. In such areas, computational models perform much worse, simply because we do not have very good criteria yet that we can use for informing them.

In summary, the opportunities that arise from engaging in computational modeling for humanists are both practical and conceptual. From the practical perspective, we can use computational models as tools to get a particular analytic task done (e.g., part-of-speech tagging) that may be a prerequisite for getting on to other, more complex kinds of analysis (e.g., analysis of genre-specific syntactic patterns). From the conceptual perspective, computational modeling can assist us in devising better models in areas we do not yet understand very well (cf. Underwood, this volume). And finally, using computational models will provoke reflection on some of our long-standing assumptions about language, pushing us to revise or revive them. In linguistics, cases in point are Chomsky's assumptions about language learning which have repeatedly been called into question (cf. Lappin and Shieber, 2007, on insights from machine learning) or the revival of the Firthian assumption of the context-dependent nature of meaning ("You shall know a word by the company it keeps," Firth, 1957, p. 11), now so popular in linguistic semantics.

Note

1 $\propto$ stands for *proportional to*; the actual probability can be derived by normalizing with the sum of the right-hand-side over all classes. For notational convenience, we conflate random variables with their values.

References

Bod, R., 2013. *A New History of Humanities. The Search for Principles and Patterns from Antiquity to the Present.* Oxford: Oxford University Press.

Bresnan, J. and Ford, M., 2010. Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English. *Language*, 86(1), pp. 186–213.

Brill, E., 1992. A Simple Rule-Based Part of Speech Tagger. In: Association for Computational Linguistics, *Proceedings of the Third Conference on Applied Natural Language Processing (ANLC '92).* Trento, Italy, March 31–April 3, 1992. Stroudsburg, PA: Association for Computational Linguistics.

Brown, P.F., deSouza, P.V., Mercer, R.L., Della Pietra, V.J., and Lai, J.C., 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4), pp. 467–79.

Bybee, J., 2010. *Language, Usage, and Cognition.* Cambridge, UK: Cambridge University Press.

Chambers, N. and Jurafsky, D., 2009. Unsupervised Learning of Narrative Schemas and their Participants. In: ACL and AFNLP (Association for Computational Linguistics and Asian Federation of Natural Language Processing), *Proceedings of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.* Singapore, August 2–7, 2009. Stroudsburg, PA: Association for Computational Linguistics.

Chater, N., Clark, A., Goldsmith, J., and Perfors, A., 2015. *Empiricism and Language Learnability.* Oxford: Oxford University Press.

Church, K., 2011. A Pendulum Swung Too Far. *Linguistic Issues in Language Technology (LiLT)*, 6(5), pp. 1–27.

Crocker, M.W., 2010. Computational Psycholinguistics. In: Clark, A., Fox, C., and Lappin, S. (Eds.) 2010. *The Handbook of Computational Linguistics and Natural Language Processing.* Oxford, UK: Wiley-Blackwell. Pp. 482–513.

Firth, J.R., 1957. *Papers in Linguistics 1934–1951.* London: Oxford University Press.

Francis, W.N. and Kučera, H., 1982. *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston, MA: Houghton Mifflin.

Goldwater, S. and Griffiths, T., 2007. A Fully Bayesian Approach to Unsupervised Part-of-Speech Tagging. In: ACL (Association for Computational Linguistics), *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics.* Prague, Czech Republic, June 23–30. Stroudsburg, PA: Association for Computational Linguistics.

Griffiths, T., Steyvers, M., Blei, D.M., and Tenenbaum, J., 2004. Integrating Topics and Syntax. In: Bottou, L., Saul, L.K., and Weiss, Y. (Eds.) 2005. *Advances in Neural Information Processing Systems 17.* Cambridge, MA: MIT Press, pp. 537–44.

Halevy, A., Norvig, P., and Pereira, F., 2009. The Unreasonable Effectiveness of Data, *IEEE Intelligent Systems*, 24(2), pp. 8–12.

Himmelmann, N., 2012. Linguistic Data Types and the Interface between Language Documentation and Description. *Language Documentation & Conservation*, 6, pp. 187–207.

Hinrichs, U., Alex, B., Clifford, J., and Quigley, A., 2014. Trading Consequences: A Case Study of Combining Text Mining & Visualisation to Facilitate Document Exploration. In: ADHO (Alliance of Digital Humanities Organizations), *Proceedings of DH 2014.* Lausanne, Switzerland, July 7–12, 2014. Available at: http://dharchive.org/paper/DH2014/Paper-373.xml.

Lappin, S. and Shieber, S.M., 2007. Machine Learning Theory and Practice as a Source of Insight into Universal Grammar. *Journal of Linguistics*, 43, pp. 1–34.

Manning, C.D. and Schütze, H., 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marcus, M., Santorini, B., and Marcinkiewicz, M.A., 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), pp. 313–30.

Markov, A.A., 1913. An Example of Statistical Investigation of the Text *Eugene Onegin* Concerning the Connection of Samples in Chains. *Proceedings of the Academy of Sciences (St. Petersburg)*, 7, pp. 153–62.

Mimno, D., Broadwell, P.M., and Tangherlini, T.R., 2014. The Telltale Hat: LDA and Classification Problems in a Large Folklore Corpus. In: ADHO (Alliance of Digital Humanities Organizations), *Proceedings of DH 2014*, Lausanne, Switzerland, July 7–12. Available at: http://dharchive.org/paper/DH2014/Paper-163.xml.

Piantadosi, S.T. and Gibson, E., 2014. Quantitative Standards for Absolute Linguistic Universals. *Cognitive Science*, 38(4), pp. 736–56.

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

Ramage, D., Manning, C.D., Dumais, S., 2011. Partially Labeled Topic Models for Interpretable Text Mining. In: ACM (Association for Computing Machinery), *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Diego, CA, August 21–24. New York: Association for Computing Machinery.

Robins, R.H., 1997. *A Short History of Linguistics*. London: Longman.

Rosenfeld, R., 2002. Two Decades of Statistical Language Modeling: Where Do We Go from Here? *Proceedings of the IEEE*, 88(8), pp. 1270–8.

Schmid, H., 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK: UMIST. Pp. 44–9.

Steyvers, M. and Griffiths, T., 2007. Probabilistic Topic Models. In: Landauer, T., McNamara, D., Dennis, D., and Kintsch, W. (Eds.) 2007. *Latent Semantic Analysis: A Road to Meaning*. Mahwah, NJ: Laurence Erlbaum. Available at: http://psiexp.ss.uci.edu/research/papers/SteyversGriffithsLSABookFormatted.pdf.

Toutanova, K. and Manning, C.D., 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In: ACL (Association for Computational Linguistics), *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*. Hong Kong, China, October 7–8. Stroudsburg, PA: Association for Computational Linguistics.