Distributional Parts of Speech

Bernhard Wälchli (University of Bern) bernhard waelchli(a)isw unibe ch

Definition: A distributional part of speech is any set of wordforms which can be obtained by clustering from a corpus on the basis of distributional information in the raw non-annotated corpus only without any resort to previous semantic, pragmatic, syntactic or morphological analysis (other than segmentation into wordforms). Distributional parts of speech are useful to the extent they happen to be congruent with semantic groups of wordforms.

1. Selected approaches to the typology of word classes

1.1 Introduction

> « Quand nous répartissons les mots en parties du discours nous procédons à peu prés comme quelqu'un qui, cherchant à résumer ce qu'ils sont des gens qui l'entourent, dirait que parmi eux il y a des bruns et des blonds, il y a des mathématiciens, des professeurs, et qu'il y a aussi des gens intelligents » (Steblin-Kaminsky quoted after Garde 1981: 156 and Nau 2001: 10, original in Russian).

> "The surface structure of this complacency is readily identifiable with the terminological vagueness seemingly endemic in this subject: familiar terms, like 'partial conversion', 'full word', 'adverb' or 'particle' have been bandied about in a cavalier way, with little attention being paid to the extent of their intelligibility" (Crystal 1967: 24).

There are many different approaches to parts of speech and these are not mutually compatible. There is not even any agreement about basic assumptions let alone the validity of arguments.

> "Recent years have seen considerable convergence in descriptive and analytic practices, including steps toward a standardized glossing system (since Lehmann 1982) and a unified ontology that must underly it. We have made less progress in standardizing the practices of argumentation, yet until we make these explicit we will be left with a situation where what counts as evidence for one linguist will be deemed irrelevant by another. This leaves our field roughly where microbiology was before Koch's postulates laid down guidelines for how a researcher demonstrates that infection by a microbe causes disease.
>
> Because the assumptions that underly argumentation are so numerous, and interact in so many ways, developing a set of convergent rules of argumentation is a huge task for the field. The very different responses from our distinguished commentators show how far we still are from having an agreed set of rules of argumentation within word class typology" (Evans & Osada 2005b: 456).

Surveys of approaches to word class typology:
Linguistic Typology 9-3: Hot debate between Evans & Osada, Peterson, Hengeveld & Rijkhoff, and Croft
Nau (2001): Chapter 2 "Was sind Wortarten und wofür braucht man sie?"
Plank (1997): Bibliography
Surveys and collections of papers: Anward et al. (1997), Evans (2000), Vogel & Comrie (2000), Ansaldo et al. (2008)

1.2 Selected approaches

- **Form-classes and distribution**: American structuralism: Harris (1944, 1951), Bloomfield (1933), Sapir (1921), Fries (1952), but also Garde (1981)
- **Propositional acts and discourse**: universal parts of speech; Croft (2000, 2005), Sapir (1921), Hopper & Traugott (1984)

- **Functional Grammar**: based on predication function; Hengeveld (1992), Rijkhoff (2002), Hengeveld & Rijkhoff (2005)
- **"Neo-classical" approach:** language-particular word classes with semantic heuristics; Evans & Osada (2005)
- **Decomposition**: different classes on several levels; Sasse (1993), Broschart (1997)
- **Word classes are part of grammatical meta-language**: Crystal (1967), Nau (2001)
- **Categorial grammar** (Gil 2000, 2008)

1.2.1 Form-classes and distribution

Parts of speech are language-particular form-classes:

> "...no logical scheme of the parts of speech—their number, nature, and necessary confines—is of the slightest interest to the linguist. Each language has its own scheme. Everything depends on the formal demarcations which it recognizes" (**Sapir** 1921: 119).

Position analysis, radically distribution-based approach without resort to semantics

> "The procedure begins by noting the environments of each morpheme and by putting in one class all those morphemes that have similar distributions. However, in many cases the complete adherence to morpheme distribution classes would lead to a relatively large number of different classes: *hotel* would be *N*, *think* would be *V*, and *take* would be in a new class *G* since it has roughly the distribution of both *hotel* and *think*. In order to avoid an unnecessary large number of classes we say that *take* is a member of both *N* and *V*. We are studying the positions, Bloomfield's 'privileges of occurrence', common to both *take* and *think*, or those common to both *take* and *hotel*" (**Harris** 1946 / 1981: 61).

English has a multipartite part of speech system (**Hockett** 1958: 225-227)
NV class: *walk, love, cure, change, air, eye, nose, beard, elbow, finger, cut, build*
AV class: *clean, dry, thin, slow, clear, busy, idle, true*
NAV class: *fancy, faint, black, yellow, blue, brown, gray, damp*

Stems of limited occurrence:

> "We use *afraid* as an adjectival predicate attribute (*He is afraid*)...We do not add *-ly*, and we do not use the word as preposed attribute to a noun, as we do most stems in classes A, N, NA, NV, AV, and NAV. *Afraid* and its kindred can hardly belong to any class but A, but they constitute a special small subclass of that class" (Hockett 1958: 228).

Fries (1952: 63-64) assumes that "the signals of all structural meanings are formal matters that can be described in terms of form and arrangement" and that "these items operate in a system".
Class 1-4 roughly corresponding to noun, verb, adjective, adverb are defined on the basis of distributional criteria only:

Class 1: Frame A *The _(s) is/was/are/were good* and some other frames.
Class 2: Frame A *(The)* CLASS-1 _ *good*; B *(The)* CLASS -1 _ *(the)* CLASS-1; C CLASS 1 _ *there*
Class 3: Frame A Class-1 Class-2 _; B (The) _ CLASS-1
Class 4: Frame A *(The)* C3 C1 C2 C3 _

Next, classes of function words are determined on the basis of what nowadays would be called prototypes.

Group A: All words for the position in which the word *the* occurs (e.g., *no, your, two, any, all*)
Group B: Prototype *may*; Group C: *not* (only one member); Group D: Prototype *very*; Group E: prototype and; Group F: Prototype *at*; Group G Prototype *do/does/did*; Group H *there* (one member only); Group I: Prototype *when*; Group J: Prototype *after*; Group K: Prototypes *well/oh/now/why*; Group L: *yes/no*; M: *look, say, listen*; N *please*; O *lets*.

Problems: It is not made explicit how the prototypes and the number of classes are determined. Languages other than English require a distinction between wordform and lexeme rather than just "words".

Criticism by Crystal (1967): Discussing the distribution of 26 English temporal nouns in 13 contexts, Crystal (1967: 53-54) finds that "even a few criteria can produce an alarming degree of complexity and overlap". He concludes that "before we can produce a set of satisfactory definitions, we need to examine the distribution of single words much more thoroughly" (Crystal 1967: 55).

Garde (1981): Word classes defined by syntactic criteria only (**dependency** grammar): five distinctive features. Based on Russian, but "avec le souci de l'universalité des définitions proposées" (155)
1. mots isolables (interjections), 2. mots connecteurs (vides), 3. mots dominants (a. verbes b. noms), 3. mots unifonctionnels (verbes et adjectives vs. mots plurifonctionnels: substantifs), 5. mots défectifs. Le 6e trait, sémantique celui-là, permet de distinguer les mots contextuels (pronoms).
Numerals are a semantically universal class like pronouns, but a syntactic class only in few languages such as Russian (Garde 1981: 184)
Problem: How can dependency relations be established without word classes?

Basic assumption: Word classes are only language-particular and based entirely on form and distribution

1.2.2 Propositional acts and discourse as basis for word classes

Sapir (1921) advocates two different theories of word classes at the same time, a distributional (above) and a propositional one.

> "There must be something to talk about and something must be said about this subject of discourse once it is selected. This distinction is of such a fundamental importance that the vast majority of languages have emphasized it by creating some sort of formal barrier between the two terms of the proposition. The subject of discourse is a noun. As the most common subject of discourse is either a person or a thing, the noun clusters about concrete concepts of that order. As the thing predicated of a subject is generally an activity in the widest sense of the word, a passage from one moment of existence to another, the form which has been set aside for the business of predicating, in other words, the verb, clusters about the concept of activity. No language wholly fails to distinguish between noun and verb, though in particular cases the nature of the distinction may be an elusive one. It is different with the other parts of speech. Not one of them is imperatively required for the life of language" (Sapir 1921: 126)

Givón (1979: 320-322): Nouns and verbs differ in **time stability** with adjectives being intermediate between stable nouns and rapidly changing verbs.

Hopper & Traugott (1984): **Discourse** imposes categoriality on nouns and verbs. The basic categories N and V are universal lexicalizations of the **prototypical discourse functions** of 'discourse-manipulable participant' and 'reported event', respectively. In non-prototypical functions nouns and verbs are decategorialized.

> "We should like to conclude, however, by suggesting that linguistic forms are in principle to be considered as LACKING CATEGORIALITY completely unless nounhood or verbhood is forced on them by their discourse functions. To the extent that forms can be said to have an a-priori existence outside of discourse, they are characterizable as ACATEGORIAL" (Hopper & Traugott 1984: 747).

According to **Dixon** (2004: 2), three word classes—nouns, verbs and adjectives—are implicit in the structure of each human language and have (a) a prototypical conceptual basis, and (b) prototypical grammatical functions. Four core semantic types are usually associated with both large and small adjective classes: DIMENSION ('big', 'small', 'long', 'tall', etc.), AGE ('new', 'young', 'old', etc.), VALUE ('good', 'bad', 'lovely', etc.), and COLOR ('black', 'white', 'red', etc.).

Croft (2005): two opposing trends: lumping (e.g., Hengeveld) vs. splitting (e.g., American structuralism)

> "Rigorous application of the distributional method would lead to a myriad of word classes, indeed, each word class would probably belong to its own word class" (Croft 2005: 434).

> "...rethinking parts of speech as restricted typological universals, not language-specific word classes" (Croft 2005: 437).

> "...for each of the propositional act constructions, one semantic class is less marked than the other two in each of the propositional act constructions" (Croft 2005: 438).

Table 1: Croft (2005: 438)

	PROPOSITIONAL ACT	PROTOTYPICALLY CORRELATED LEXICAL SEMANTIC CLASS
a.	reference	objects (nonrelational, stative, inherent, nongradable)
b.	predication	actions (relational, dynamic, transitory, nongradable)
c.	modification	properties (relational, stative, inherent, gradable)

For Croft (2001), typology is a theory, not a method. The universal semantic domains of functional linguistics (Givón 1981, Stassen 1985, 1997, Miestamo 2005, 2007) are considered to be basic units of the theory.

Disagreement about the status of adjectives: Discourse/conceptually based for Croft (2005) and Dixon (2004), but not for Sapir (1921), Hopper & Traugott (1984) and Dixon (1977).

Basic assumption: Parts of speech are not grammatical categories of particular languages, but rather basic discourse or conceptual units.

1.2.3 Functional Grammar (Hengeveld, Rijkhoff)

Distinction between lexical units (noun, verb) and syntactic units (term phrase, predicate phrase).
Only classes of lexemes, i.e. predicates are considered (verbs, nouns, adjectives, and adverbs).
Differences between predicates are defined "in terms of the prototypical functions they fulfil in the construction of predication" (Dik 1989: 162). These functions may be assumed to be universally recognizable. Predicative use of classes cannot be taken as a criterion for the definition of parts of speech (Hengeveld 1992: 48). In order to find out whether a predicate is an adjective, its attributive use should be studied (ibid. 47).

Table 2: Types of parts of speech according to Hengeveld (1992)

		Head of predicate phrase	Head of referential phrase	Modifier of head of referential phrase	Modifier of head of predicate phrase	Languages
Flexible PoS systems	Type 1	contentive				Tongan, Samoan, Mundari
	Type 2	verb	non-verb			Quechua (Imbabura), Tagalog
	Type 3	verb	noun	modifier		Dutch
Rigid PoS systems	Type 4	verb	noun	adjective	adverb	English, Hungarian
	Type 5	verb	noun	adjective		Wambon
	Type 6	verb	noun			!Xũ
	Type 7	verb				Tuscarora

Basic assumption: parts of speech are based on universal functions which are not universally represented (often realized indirectly).

1.2.4 Language-particular word classes with semantic heuristics ("neo-classical")

> "...our arguments are neo-classical in the sense that we wish to retain the strengths of the structuralist tradition" (Evans & Osada 2005b: 450).

Three criteria for establishing lack of word class distinctions:
Compositionality: any semantic differences between the uses of a putative "fluid" lexeme in two syntactic positions (say argument and predicate) must be attributable to the function of that position.
Bidirectionality: it is not enough for Xs to be usable as Ys without modification: it must also be the case that Ys are usable as Xs.
Exhaustiveness: "...it is not sufficient to find a few choice examples which suggest word class flexibility. Since word classes are partitionings of the entire lexicon, equivalent statements need to hold for all relevant words in the lexicon that are claimed to have the same class" (Evans & Osada 2005: 378).

Mundari makes wide use of zero conversion, resulting in frequent **heterosemy** (Lichtenberk 1991; the use of identical forms with different combinatorics and different meanings. In a sample of 3'824 lexemes, 20% are nouns only, 28% verbs only and 52% are both nouns and verbs (Evans & Osada 2005: 357, 383).

> "Introducing semantics into our heuristics allows to shave off a major cause of apparent distributional chaos, which results from the differential effects of polysemy on the distribution of words" (Evans & Osada 2005b: 451).

Basic assumption: Word classes are entities in the description of individual languages, but semantic criteria are indispensable.

1.2.5 Decomposition

Lexical categories vs. syntactic categories

> "...even if it were true that there is a very close correlation between traditional nouns and verbs and particular syntactic categories...this would not mean that languages with lexical classes other than nouns and verbs are inconceivable, whose syntactic categories...are defined independently of the lexical classes in question" (Broschart 1997: 130).

> Samoan: "Many, perhaps the majority of, roots can be found in the function of verb phrase and noun phrase nuclei and are, accordingly, classified as nouns and verbs...This does not mean that a noun can be used as a verb or a verb as a noun or that we have two homophonous words...Rather, it means that in Samoan the categorization of words into nouns and verbs is not given a priori in the lexicon" (Mosel & Hovdhaugen 1992: 76).

> „Viele Linguisten arbeiten heute implizit oder explizit mit einem mindestens vier Ebenen definierten Begriff von lexikalischer Kategorie: formal-morphologisch, semantisch (ontologisch), syntaktisch, und diskurspragmatisch (Referenz und Prädikation). Das ist eine sehr starke Annahme bezüglich einer ganz spezifischen Merkmalbündelung" (Sasse 1993: 192).

1. Formal parameter (inflection, derivation, distribution)
2. Syntactic parameter (slot-filler relationship)
3. Ontological-semantic parameter
4. Discourse-pragmatic parameter (reference, predication, modification) (Sasse 1993: 196)

> „Eine andere Möglichkeit besteht in der überlappenden Distribution der formalen Mittel, wie dies etwa im Ungarischen oder Türkischen der Fall ist...So sind z.B. die Personalendungen des ungarischen Verbs weitgehend identisch mit den Possessivsuffixen des Nomens. Da jedoch das ungarische Verb insgesamt ein ganz anderes Flexionspotential aufweist als das Nomen, kann von einer mangelnden kategoriellen Distinktion keine Rede sein; die Endungen sind nicht etwa vage in Bezug auf die Unterscheidung von Possession und Subjektkongruenz, sondern als homophone Formen anzusehen, die dem Ausdruck unterschiedlicher grammatischer Kategorien dienen. Ein solcher Fall ist durchaus zu unterscheiden von Fällen totaler Identität von Possessiv- und Personalaffixen, wie er etwa in einigen Indianersprachen vorliegt“ (Sasse 1993: 197, based on Walter 1981).

Type/token languages (Tongan) vs. noun/verb languages (Latin/German) (Broschart 1997: 157)

TOKENS	[+ref]	(a) tense-marked	or (b)	article-marked
TYPES	[-ref]	(a) not tense-marked	or (b)	not article-marked
NOMINAL	[-pred]	(a) article/gender marked	and (b)	not tense marked
VERBAL	[+pred]	(a) tense-marked	and (b)	not article/gender marked

Basic assumption: Word classes are language-particular. Syntactic categories and lexical categories must be distinguished.

1.2.6 Grammatical meta-language, "usefullness" of word classes

Usefulness to the linguist or teacher: classes have to be few and fairly general and have some degree of intuitive coherence (**Crystal** 1967: 41)

> "It is frequently assumed that one can satisfactorily describe the word classes of (say) English before going to the 'meaty' part of grammar, for which the classes are seen merely as a kind of grammatical shorthand. This is complacency, because to isolate word classes in such a way is both misleading and distorting: word classes should not be taken as being in some way part of a terminological preamble to grammar, because in a real sense they assume a grammar before one can begin to talk about them. Their definition is an abstraction from grammatical and other criteria – not directly from data – and their purpose is ultimately to act as the constituents of a grammatical meta-language, which one manipulates to display more interesting syntactic relations." (Crystal 1967: 25)

Many classes of word classes for different purposes (**Nau** 2001)
Classifications of PoS are based implicitly on a series of ideal postulates (P.1-P.10) (Nau 2001: 8-10)
P.1: **Exhaustivity**: Every word can be assigned to a word class.
P.2: **Unequivocality**: Every word belongs to exactly one class (except homonyms).
P.3: **Taxonomy**: The classes can be ordered hierarchically in a taxonomic system.
P.4: **Uniqueness**: Each language has only a single word class system.
P.5: **Intuition**: Classification is easy. With the possible exception of a few difficult cases, the word class of any word can be determined easily even by laymen.
P.6: **Few classes**: The number of classes is small.
P.7: **Clear boundaries**: The classes are neatly delimited.
P.8: **Few criteria**: The classes are determined by a small number of simple criteria. Usually two or three criteria are assumed.
P.9: **Consistency**: All classes must be determined by the same criteria or at least by criteria of the same kind.
P.10: **Form-classes**: Word classes are syntactic categories.

The "word class trap": The explanandum becomes the axiom.

> "Each grouping of words making sense ["sinnvoll"] can be called word class. Whether a grouping makes sense, can be decided only with regard to an exactly determined purpose" (Nau 2001: 23).

Different purposes/applications: reference grammars, dictionaries, linguistic theory, cross-linguistic comparison

Classes can be formed according to different criteria:

- Lexeme classes, wordform classes, functional classes
- Semantic, morphological, syntactic classes
- Classes formed on the basis of one or several criteria
- Given classes (top-bottom) vs. obtained classes (bottom-top)

Table 3: Classes of lexemes vs. classes of wordforms vs. functional classes (Nau 2001: 24)

Lexem class	L-1, lexeme SCHNELL		
Wordform class	Wf-1, inflected forms *schnelle, schnelles, schnellem...*	Wf-2, non-inflected base form *schnell*	
Functional class	F-1, attributive *der schnelle Igel*	F-2, predicative *der Igel ist schnell*	F-3, adverbial *der Igel rannte schnell*

Given vs. obtained classes (*Gegebene vs. gewonnene Klassen*): Difference in method of class formation
Given classes (top-bottom): we assume that there are “nouns” and look for features that distinguish them
Obtained classes (bottom-top): we first take words and consider their features and form groups of them without any hypothesis what classes should be like.

> Similarly, Crystal: Establishing vs. describing word classes: “The problem of setting-up word classes is basically a question of discovery procedures, and the issue arising here are very different from the purely descriptive problem, where word class criteria are verified against an independently-verifiable grammar.” (Crystal 1967: 25)

> “In my view wordclasses are units of linguistics and not units of language” (Nau 2001: 29).

Basic assumption: There are very many word classes. They are formed according to various criteria and for different purposes. Word classes are units of linguistics, not of language.

1.2.7 Categorial grammar

Gil (2008 and elsewhere) proposes a **categorial grammar** approach to word classes. Word class categories are purely syntactic and both words and constituents of words are assigned the same kind of category labels. Every language has at least the category S^0, which stands for any word or constituent that can occur as a complete non-elliptical sentence. Further categories can be derived by two operators (slash and kernel).

According to Gil (2008), Jakarta Indonesian has two word classes: an open syntactic class S^0 and a small heterogeneous closed category S^0/ S^0. Almost all words in Jakarta Indonesian are S^0 and exhibit identical syntactic behavior. There are no syntactic differences between words referring to things and words referring to activities. S^0 words can occur as complete non-elliptical sentences and can combine with any other S^0 word (sometimes semantically anomalous) (Gil 2008: 653). In addition there is a small number of preceding and following S^0/ S^0 words, which cannot occur without an S^0 word, such as (preceding) *ke* ‘to’, *yang* REL.

Basic assumption: Word classes follow from syntactic segmentation. More complex categories are formed from simpler ones by means of category formation operators.

2. Avoidance of word classes in typology

> "For the purposes of this map, these distinctions in word class are ignored: a word is treated as an adjective, regardless of its word class in the language, as long as it denotes a descriptive property. The map also ignores the question of whether the adjectives are modifying nouns directly or whether they are the predicate of a relative clause which is modifying the noun" (Dryer 2005).

Also other WALS typologies avoid word classes:

> "...is it really an adjective, or is it a noun, or perhaps a member of some other, less differentiated part of speech? In this chapter, we have been using the terms noun and adjective in the traditional manner, as labels for semantic categories rather than syntactic ones" (Gil 2005).

No explicit reference to parts of speech or word class is made even in Stassen (2005), who discusses the "verbal or nonverbal status of any case of predicative adjective encoding in a language-independent fashion."

> "The basic distinction is between those languages in which predicative adjectives are encoded in a way that is parallel to predicative verbs, and those languages in which the encoding of predicative adjectives and of verbs is different" (Stassen 2005).

According to (Croft 2005: 434) Stassen (1997) applies the same approach as Croft and it is true that Stassen applies a universal functional domain approach. However, Stassen (1997) does explicitly not deal with word classes: "It is of the utmost importance to keep in mind that this book is not an essay in the universal theory of PARTS OF SPEECH; it merely describes the typological characteristics of various lexical categories with respect to predicative encoding" (Stassen 1997: 30).
Stassen explicitly favors the avoidance strategy. He considers the part of speech discussion to be arbitrary, since there is no successful proposal for ranking criteria: "Given that there is no objective manner to weigh these conflicting criteria against one another, the question of assigning word class status to certain lexical items in a language may sometimes result in indeterminacy or arbitrariness" (Stassen 1997: 31).

-> Obviously, it is doubtful whether word classes are a necessary unit of typology: many typologists avoid referring to word classes when treating phenomena traditionally associated with word classes. One gets the impression that avoiding word classes makes typologies more efficient and more empirical.

-> Linguistic typology has no uniform attitude toward parts of speech. On the one hand, it is clear that it is one of the most fundamental issues for typology. The first chapter in volume 1 of Shopen (ed.) (1985) *Language Typology and Syntactic Description* treats parts-of-speech systems (Schachter 1985). A large part of *Linguistic Typology* 1 (1997) is devoted to the discussion of word classes. On the other hand, there is no agreement about the most basic assumptions and typologies are most efficient if they avoid word classes.

3. An algorithmic approach to parts of speech and word classes

3.1 Basic elements of an algorithmic approach to parts of speech and word classes

3.1.1 Parts of speech vs. word classes

Most approaches use part of speech and word class as synonyms (but not Croft 2005: parts of speech are universal and word classes are language-specific).

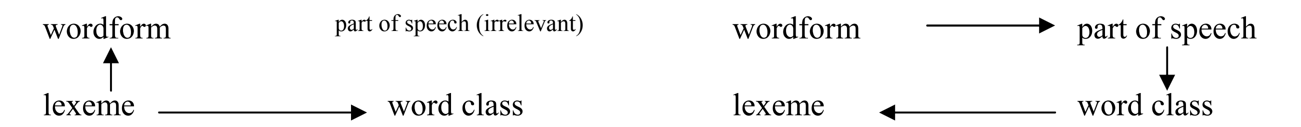
Here:	PARTS OF SPEECH:	CLASSES OF WORDFORMS
	WORD CLASSES:	CLASSES OF LEXEMES

The possibility of studying classes of wordforms does not occur to most non-computational linguists, an exception is Nau (2001). However, classes of wordforms are typical objects of study in computational approaches and corpus linguistics usually without discussion of theoretical implications, see e.g. Biemann (2006b).

Why should typologists be interested in classes of wordforms? Aren't these completely uninteresting?

The reason is that it cannot be avoided. Parts of speech (classes of wordforms) are more basic entities than word classes. They can be detected directly in corpora without any previous analysis of text except segmentation into wordforms. I will argue here that word classes, paradigms, and lexemes can be obtained only if parts of speech have been obtained already.

Traditional processing chain	Processing chain in unsupervised algorithmic word class analysis



3.1.2 Partitioning

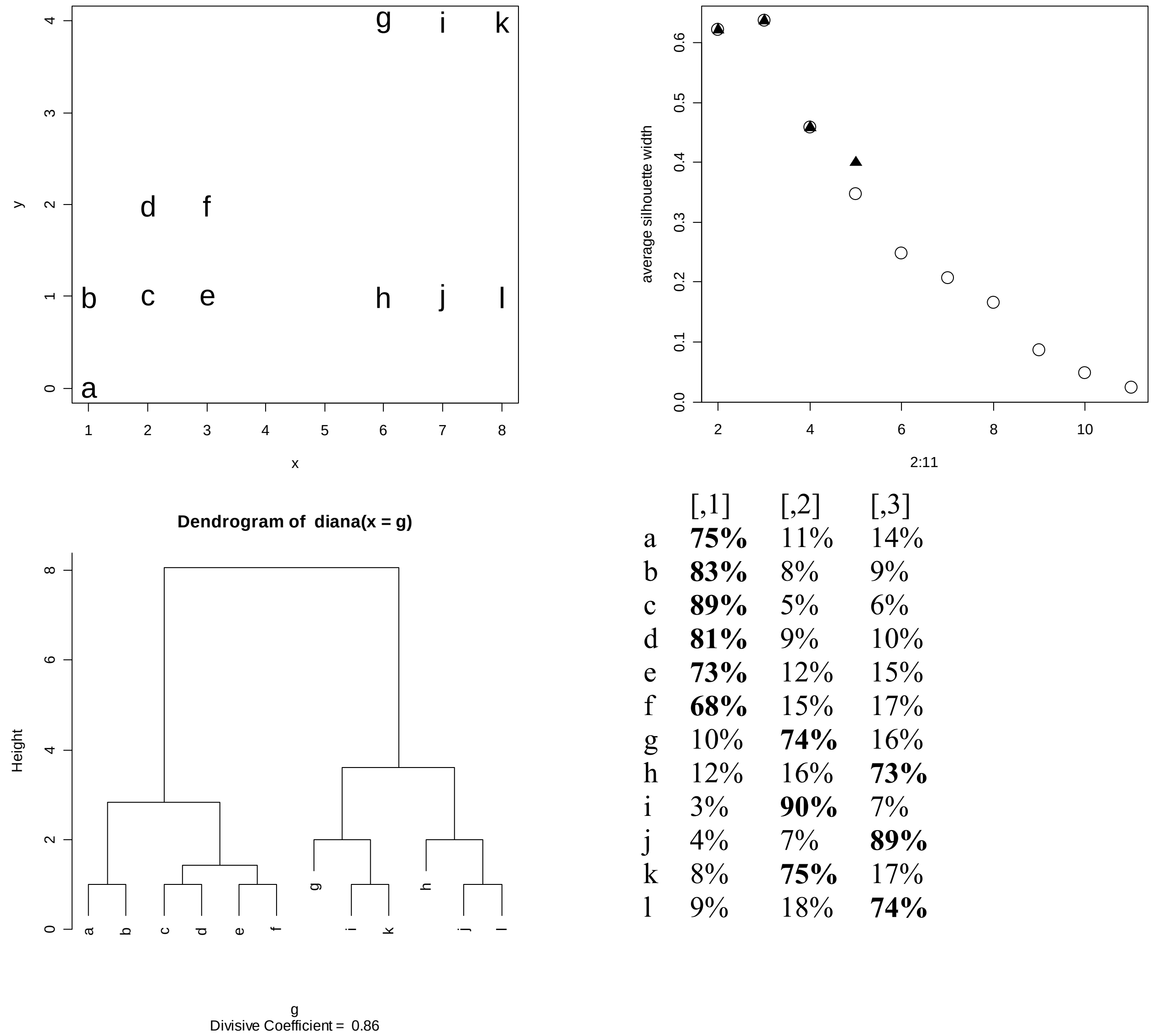
Assigning forms to word classes is partitioning. This term occurs as non-technical term in the literature (e.g., Evans & Osada 2005: 378), but it is a technical term in data mining, more specifically cluster analysis. Every linguist dealing with word classes should have some idea what partitioning is when understood as a technical term.

"Cluster analysis is the art of finding groups in data" (Kaufman & Rousseeuw 2005: 1), this is exactly what finding word classes is about. Unfortunately, partitioning is not as simple as addition where there is only a single method. In recent years, hundreds of slightly different partitioning methods have been developed. Let us consider three examples from Kaufman & Rousseeuw (2005) which are implemented in the free software tool R: `pam()`, `diana()`, and `fanny()`

pam(x,k) is Partitioning of the data x into k clusters Around Medoids. The user has to choose how many clusters there should be. For the data in Figure 1 converted into a distance matrix with `daisy()`, for k=2 the clusters are {a,b,c,d,e,f} {g,h,i,j,k,l} and for k=3 the clusters are {a,b,c,d,e,f} {g,i,k} {h,j,l}. For k=2 the medoids (in linguistic terms, the most prototypical member of the cluster) are c and i and for k=3, c, i, j. While the number of clusters must be prespecified in `pam()`, there is a measure for the goodness of the partitioning which shows that 3 clusters is slightly better than 2 clusters and that the goodness decreases rapidly from 4 to 11 clusters (Figure 1, top-right circles). A linguistic interpretation of "medoid" would be the most prototypical example of a word class.

Partitioning methods can directly be compared to Nau's (2001) ideal postulates. `pam()` meets exhaustivity, unequivocality, few criteria (x,y axes values), but not taxonomy. The criterion "Few classes" is up to the user, s/he can choose the number of clusters. There is no uniqueness, but various results of clustering can be evaluated, so that we can see that 3 clusters is best, but 2 clusters are about equally good. Other clustering methods, such as `diana()` [Divisive Analysis Clustering], are hierarchical (Figure 1,bottom-left). Others, such as `fanny()` [Fuzzy Analysis Clustering] assign elements to clusters to a certain extent. The best partitioning with `fanny()` is again with 3 clusters and the percentage to which the elements belong to the three clusters is given in Figure 1 bottom-right. This partitioning method fails the unequivocality ideal, but unequivocality can be forced (closest hard clustering) if the cluster with the highest percentage is chosen (bold) which yields the same result as with `pam()` in this case (but not necessarily with all datasets). The goodness values for `fanny()` are similar to those of pam (Figure 1, top-right, triangles) in this dataset.

Figure 1: Partitioning a dataset with `pam()`, `diana()`, and `fanny()`
(top-left: elements to be clustered; top-right: goodness of partition (average silhouette width); bottom-left: hierarchical clustering; bottom-right: fuzzy analysis clustering with three clusters)



	[,1]	[,2]	[,3]
a	**75%**	11%	14%
b	**83%**	8%	9%
c	**89%**	5%	6%
d	**81%**	9%	10%
e	**73%**	12%	15%
f	**68%**	15%	17%
g	10%	**74%**	16%
h	12%	16%	**73%**
i	3%	**90%**	7%
j	4%	7%	**89%**
k	8%	**75%**	17%
l	9%	18%	**74%**

Linguists should be familiar with a technical understanding of the notion of partitioning because there is a wide-spread misunderstanding that distributional criteria necessarily lead to a “myriad of word classes” (Croft). This argument is ascribed to the American Structuralism: the possibility to use distribution as single criterion is rejected by beating the American Structuralism with their “own” argument: “Form-classes are not mutually exclusive, but cross each other and overlap and are included one within the other, and so on...For this reason, a system of parts of speech in a language like English cannot be set up in any fully satisfactory way: our list of parts of speech will depend upon which functions we take to be the most important” (Bloomfield 1933: 269). However, it follows from the nature of partitioning that the modern linguists’ argument is wrong. The number of clusters resulting does not depend on the distribution, but on the method of partitioning applied. Bloomfield is perfectly right in claiming that there will always be several possibilities to form clusters; this is an unalterable property of cluster analysis. If there is a need for a small number of classes there are many clustering techniques which yield small numbers of classes from every dataset. And even though there are a large number of clustering techniques, many of them are rigorous and replicable (however, it is sometimes hold that clustering techniques containing a random element are more powerful).

Partitioning is the “statistical rationale” postulated by Crystal (1967: 47): “A statistical rationale of the criteria for word classification seems to be the only alternative to the unqualified arbitrariness which Bloomfield stated was implicit in the definition of English word classes”

3.1.3 Distribution and semantics

Co-occurrence statistics: many syntactic and semantic relationships between wordforms in texts can be detected by statistical methods. In order to make use of this rich source of information one need not be a radical behaviorist assuming that semantics and syntax *are* distribution. But it cannot be denied that many aspects of semantics and syntax are *reflected* in corpora.

> “The goal of co-occurrence statistics is to extract pairs of words {i.e., wordforms, BW} that are associated from a corpus. The underlying assumption is that while generating a text, people are complying to syntactic and semantic restrictions of their (natural) language in order to produce correct sentences. When analyzing a large quantity of text (a text corpus), words that tend to appear together will reflect these linguistic restrictions. While it is generally possible to produce sentences containing arbitrary pairs of words, in most of the cases the words appearing together will have something to do with each other and statistics will be able to cut out the noise.
>
> The joint occurrence of words within a well-defined unit of information, for example, a sentence, a whole document, or a word window, is called co-occurrence” (Cysouw, Biemann & Ongyerth 2007: 160-161).

“...two kinds of dependencies in a corpus:

1. **Syntagmatic**: language units representing compliance due to an assumed attribute, such as words or morphemes which either attract or inhibit cooccurrences. ...the word *ich* attracts the morpheme +*e* in the corresponding verb and inhibits the morpheme +*st*.
2. **Paradigmatic**: language units representing the assumed attribute are not easily interchangeable, despite belonging to the same paradigmatic class. Furthermore, various representations of an attribute belonging to the same paradigmatic class are mutually exclusive, implying that co-occurrences of these, such as the direct co-occurrence of *Ich* with *Du*, are far less probable and mainly confined to special language usages” (Bordag 2007: 56).

Virtual vs. actual distributions

The whole typological and descriptive literature (except Harris) assumes that distributions are modal (or virtual) rather than phenomenological (or actual). For instance, Schachter & Otanes (1972) write about Tagalog that nouns and verbs cannot be distinguished on the basis of distribution: “there is virtually no context in which a noun occurs in which it cannot be replaced by a verb or verb phrase” (1972: 65). However, it is simply not the case that verb forms and noun forms have the same distribution in actual corpora in Tagalog. Wherever individual forms sort in, nouns and verbs very clearly emerge in different clusters in Tagalog in automatic part of speech analysis (see 3.3.1 below). Put differently, even though Tagalog nouns and verbs may have the potential of having the same distribution virtually, they make very little use of this potential in an actual corpus. The same holds for Indonesian. The huge macro-class S^0 may be virtually identical in syntax due to the same distributional privileges (Gil 2008: 639), but nouns and verbs do not exhibit identical behavior in actual distribution in texts.

> “The distribution of an element is the total of all environments in which it occurs” (Harris 1951 / 1960: 15-16)

NOT: in which it can occur

A virtual/modal approach to distribution seems to be restricted to linguistics. It statistics it is common to study actual distributions.

3.1.4 The Large-Number-of-Rare-Events (LNRE) problem in quantitative linguistics

That languages have only one word class or that every word is a class of its own are no options because of the LNRE problem. (It does not follow from this that languages with a verb=noun class are impossible, but natural languages with a single word class are impossible.)

> "...word frequency distributions, and even more so the distributions of bigrams and trigrams, are characterized by large numbers of very low probability elements. Such distributions are referred to as LNRE distributions, where the acronym LNRE stands for Large Number of Rare Events (Chitashvili and Khmaladze, 1989; Baayen, 2001). Many of the rare events in the population do not occur in a given sample, even when that sample is large. The joint probability of unseen words is usually so substantial that the relative frequencies in the sample become inaccurate estimates of the real probabilities" (Baayen 2008: 229).

A system of constructions (or syntax) can never emerge if the many rare events are not grouped into classes. (It may be the case that all "constructions" in early L1 acquisition are formulas, but as soon as the number of wordforms starts expanding rapidly, these must somehow be organized into clusters.) The LNRE nature of language rules out the possibility suggested by Nau (2001: 29) that word classes might be only entities of linguistics and not of language. However, it does not follow from the LNRE nature of language that word class systems must be unique or that word classes must have clear boundaries. Neither does it follow from it whether the classes are mainly syntactic, mainly semantic, or mainly morphological. It only follows that there must be a number of clusters that are not rare events and that they form constructions which are no rare events. Whether the constructions are more basic (as claimed by Croft) or the elements cannot be decided on the basis of the LNRE nature of language.

The LNRE nature of language is also a problem for Evans & Osada's (2005) exhaustiveness criterion. It is very likely that word class systems are organized such that the wordforms occurring in a large corpus are enough for establishing at least the major word classes.

The first and foremost purpose of word classes is to make the LNRE phenomenon language manageable. There must be clusters of wordforms because most of the individual wordforms themselves are too rare. Rare words, however, are indispensable for language as an effective tool for communication because rare words convey more information than frequent words.

3.1.5 Unsupervised learning and processing chains

"Unsupervised learning" means assuming an empiricist or "Martian" perspective. We have a set of utterances or corpus and the task is to find a procedure by which we can decode the structure of any language without knowing anything about particular languages. The term "learning" is a bit misleading; it does not imply that children actually learn languages that way. However, learning is an algorithm and algorithms are basically independent of the medium where they are implemented. Language learning algorithms can be modeled in computers. Given sufficiently large corpora of natural languages it is possible to extract certain structures from these corpora if we can come up with the right algorithm and proceed from basic to more complex structures along the right kind of processing chain.

Particularly important is that processing chains are not allowed to be circular. For instance, we cannot define word classes on the basis of "tense" or "case" if these categories have not been previously identified by an algorithm which does not make reference to word classes. Here it is assumed that word classes are very basic entities. They are needed before we can start doing anything like syntax. This is implicitly assumed in many approaches where word classes are given "separate discussion towards the beginning", a practice criticized by Crystal (1967: 25). If word classes are categories of the beginning, the consequence is that only entities which are given before word classes have been abstracted may be used for defining them, a condition which is usually violated in the discussion of word classes.

However, it is a common technique in uncovering complex algorithms to start in the middle (see Dennett 1991: 61). Here, we start at a stage where wordforms have been segmented (at the level of analysis writing offers). We can make reference to any form of distribution of wordforms in a corpus, but not to any structure beyond that (no dependency or constituents or morphemes or grammatical categories).

The approach is bottom-top: there are no given classes, all classes are obtained (see 1.2.6).

3.1.6 An evolutionary perspective

There might be a universal language learning algorithm, by which the structure of any language can be acquired from a corpus of (contextually embedded) utterances. The same universal process may generate very different structures when exposed to different input.

This does not mean necessarily that every child learns a language exactly the same way (there is much evidence that this is not the case); it does not even mean that all speakers of a language or dialect necessarily have the same mental representation. Rather it means that natural languages are constrained in particular ways which distinguishes them from many other conceivable languages which do not happen to exist because they cannot be replicated. Whatever algorithm is applied, it must lead to the same or a very similar result in the end (and only in the end) for any particular language. It is well possible that there are various possible processing chains that lead to the same result.

All natural languages have **descent**. Put differently, natural languages are **replicated** and thus a possible human natural language is a language that can be replicated within a group of human beings (within a speech community).

Croft (2000) assumes that the **utterance is the basic unit of replication**. According to Croft’s theory of utterance selection for language change, a language is the population of utterances in a speech community (rather than a system of contrasts of signs or idealized speaker/hearer competence). The utterance embodies linguistic structure and is the carrier of linguistic replicators (linguemes). Language change evolves through altered replication in contrast to normal replication where the replicated structure is identical with the parent structure.

There is no direct pathway from mental representation to mental representation in replication. Ergo, **all non-universal elements of linguistic structure must be fully contained in the utterance (in whatever form)**. A sufficiently large corpus of a natural language must contain all necessary information to build a fully productive representation of that language if the right universal algorithm is applied to it (see Mayer & Wälchli 2007).

3.1.7 The principle of congruence

Form and meaning happen to be congruent in languages with descent. In order to be replicable, form and meaning must match. This is why a considerable component of linguistic structure can be extracted from form alone (from a corpus).

Sasse and Nau (see 1.2.5-6 above) are right that it cannot be taken for granted that features bundle on various levels. However, it seems to be a design-property of natural languages that different levels happen to be congruent to a large extent. Put differently, items can be classified on the basis of criteria on one level and then it is often the case that there is a partial or full correspondence of obtained categories with another level.

Instances of congruence are manifold: between form and meaning, between lexical categories and syntactic categories, between syntax and morphology.

Clusters of wordforms obtained from purely distributional criteria often tend to be semantic classes, lexical form-classes, and/or syntactic classes or at least elements of lexical or syntactic form-classes. Distributional parts of speech are often building blocks of paradigms.

3.1.8 Procedural universals vs. structural universals

Universals of parts of speech and word classes are procedural universals and not structural universals. The same or similar learning algorithms are applied to very different input which results in very different structure and mental representation. Universals of parts of speech and word classes are the fully explicit processing chain how parts of speech and word classes can be identified in a corpus.

The same program is used to perform automatic part of speech clustering on the basis of a large corpus resulting in very different results across different languages.

3.1.9 Emerging categories

> "If they are neither completely in the genes nor completely in the linguistic input, where, then, do grammatical complexities come from? It seems ineluctable to assume that it is the interaction between the two that shapes the end product. To take that conclusion one step further, we have strong reasons to consider the possibility that this interaction produces a result that is not merely the sum of its component parts but is more. The interaction is a source of novelty and complexity. In other words, it generates quantitatively new phenomena, *emergents*, whose complexity is not explicitly preformed, but arises as an automatic consequence of the interaction, that is by *self-organization*, and goes beyond that found either in the initial conditions or in the input" (Lindblom 1992: 133).

The universal algorithm which first learns the parts of speech and derives word classes from them from a corpus in any language is here assumed to be very simple. It considers distribution and makes partitionings on the basis of the distribution. The structure that it generates exceeds by far in size its own size.

3.1.10 Primary-data typology and data reduction

Primary data typology is a cover term for all typological data collection processes based on primary sources rather than descriptions and on exemplars rather than abstractions. Data sources of primary data typologies are, e.g., translational questionnaires, retold stories, original texts, and parallel texts. Primary data typology has the advantage that typologies with less data reduction can be made.

Most typologies such as represented in large typological databases (e.g. WALS) have undergone extremely strong **data reduction**. There are different motivations for data reduction in typology. First of all, practical reasons: data is presented in grammars and dictionaries in a reduced form which is why a high amount of data reduction is there already before typology comes into play. Since grammars differ in the amount of detail listed about relevant domains, a common strategy is to make the typology so general that most grammars can be expected to contain the relevant information.

Another common but much more doubtful motivation for data reduction is to keep the error rate small. However, data reduction does not reduce the magnitude of the sum of errors:

> "Many researchers make the mistake of assuming that categorizing a continuous variable will result in less measurement error. This is a false assumption, for if a subject is placed in the wrong interval this will be as much as a 100% error. Thus the magnitude of the error multiplied by the probability of an error is no better with categories" (Harrell 2001: 6).

While reference grammar typology usually takes the notion of homogeneous particular languages as given much as generative grammarians conjecture the competence of ideal speaker-hearers and structuralists postulate the *langue* of speech communities, primary data typology is compatible with the idea that languages are only abstractions of populations of very similar idiolects, such as has long been claimed by the philosopher of language Fritz Mauthner:

> „In Wirklichkeit ist auch der Begriff der Einzelsprache nur ein Abstraktum für die Fülle von Ähnlichkeiten, von allerdings sehr großen Ähnlichkeiten, welche die Individualsprachen einer Menschengruppe bieten...Aber auch die Ungleichheit einer Individualsprache in verschiedenen Lebensperioden ist größer als man wohl glauben möchte..." (Mauthner 1923: 6-7).
> ["The notion of particular language is in fact only an abstraction for the mass of similarities, of admittedly very strong similarities, among the languages of individuals of a human population...But even the dissimilarity of an individual's language across different periods of life is greater than commonly believed", translation BW]

3.1.11 Typologies of doculects rather than of languages

Doculect: any documented language variety, be it as raw data (e.g., sound file), primary data (e.g. a [transcribed] text), or secondary data (description, e.g., a grammar) of whatever size. Doculect is related to language as sample to population in statistics. A doculect can thus be more or less representative of a language. The term doculect has been coined by Michael Cysouw, Jeff Good and Martin Haspelmath in

2006 at the Max Planck Institute for Evolutionary Anthropology and is first mentioned in the published literature in Bowern (2008: 8).

High levels of data reduction suggest that languages can be an object of cross-linguistic comparison. Actually, only doculects can be objects of study. We can speak of "languages" only if we are certain that different doculects of the same language would always yield the same result.

3.1.12 The psychological-reality argument

It is not shown here that distributional parts of speech are psychologically real. Thus, Dixon (1977) might be right about the priority of semantics:

> "Suppose that the item is a verb; then in order to work out which types of object noun phrase complements, say, it could occur with, the speaker would just have to keep his ears open. After a year or so he might subconsciously muse 'I have heard the verb used with THAT complements but never with FOR-TO or with POSS-ING complements' and would thus mark the item '+THAT, –FOR –TO, –POSS-ING' in his mental lexicon. Only then would he be able to use the verb productively and correctly. Obviously, this bears little relation to what happens when a speaker learns a new word, demonstrating the untenability of the 'syntax prior' position" (Dixon 1977: 24-25).

What is demonstrated here is that the information to classify word forms into parts of speech is contained in corpora without any resort to semantics and that parts of speech can be extracted from corpora. It is not shown that there is no other way to reach the goal by making use of meaning. In the spirit of the congruence principle it can be assumed that there may exist several pathways to reach the same goal. However, the distributional pathway is very convenient for typological purposes because it is fully explicit: The same process can be applied to all language corpora considered.

3.2 Automatic part of speech partitioning

Unsupervised Part-of-Speech Tagging Employing Efficient Graph Clustering (Biemann 2006b)
Parts of speech are extracted independently in two partitions:
Partition 1: wordforms with high and medium frequency
Partition 2: wordforms with medium and low frequency
The two partitions are combined. The text is tagged with the obtained classes

- Partition 1: The 10'000 most frequent wordforms are target forms (forms to be classified), they are clustered on the basis of co-occurrence with the 200 most frequent forms. A graph is constructed from context statistics [cosine similarity of two vectors w=1/(1-cos(a,b))] with a similarity threshold that removes ambiguous wordforms (homonyms). "Chinese Whispers" (Biemann 2006a) is used as partitioning method.
- Partition 2: All wordforms below rank 2000. Similarity scores between pairs of wordforms are calculated according to how many neighbours they share with log-likelihood statistics. Again, "Chinese Whispers" is used as partitioning method.
- A graph containing the clusters of both partitionings as nodes is constructed which is again partitioned by "Chinese Whispers". This results in fewer clusters.
- A lexicon is constructed from the merged partitionings, which contains one possible tag per word. Forms belonging to more than one class (homonyms) are ignored A Trigram Veterbi Tagger is used to tag the corpus. Morphologically motivated add-ons are used to guess a more appropriate category distribution based on a word's suffix.
- Biemann (2006a) applies this method to large corpora in three languages: English, German, and Finnish.

> "[Chinese Whispers] is a very basic – yet effective – algorithm to partition the nodes of weighted, undirected graphs. It is motivated by the eponymous children's game, where children whisper words to each other. While the game's goal is to arrive at some funny derivative of the original message by

passing it through several noisy channels, the CW algorithm aims at finding groups of nodes that broadcast the same message to their neighbors. It can be viewed as a simulation of an agent-based social network" (Biemann 2006a: 74).

Table 4: Tagging example (Biemann 2006b: 12)

Wordform	Cluster members (size n)
I	*I* (1)
saw	Past tense verbs (3818)
a	*a, a, the* (3)
man	Nouns (17418)
with	Prepositions (143)
a	*a, a, the* (3)
saw	Nouns (17418)
.	*. ! ?* (3)

3.3 Toward a typology of distributional parts of speech

An approach similar to the one of Biemann (2006b), but much simpler, is applied here to corpora in 23 languages of 0.6-1.5 m wordforms in size. Only frequent word forms are clustered.

The algorithm implemented in a Python program first considers the environment of all wordforms to be clustered, calculates a distance matrix out of it, which is partitioned by `pam()`, see 3.1.2. Further details will be given in Wälchli (in prep.).

The same algorithm is applied to large corpora of various languages (~1 m words). This can only be done with a convenience sample. The sample consists of doculects rather than of languages. The result depends on the concrete corpus. If another corpus of the same language is chosen, the result will be different (but similar).

Doculects considered: Albanian (Bible), Bulgarian (http://www.bultreebank.org/Resources.html), English (Ch. Dickens, M. Twain and other novels), Estonian (http://www.cl.ut.ee/korpused/baaskorpus/), Finnish (Bible, novels from www.gutenberg.org), French (V. Hugo, *Les Misérables*), German (Mommsen, *Römische Geschichte*), Modern Greek (Katharevousa, Bible), West Greenlandic (Bible), Haitian Creole (Bible), Hungarian (Bible), Indonesian (Bible), Italian (novels from www.gutenberg.org), Latin (*Vulgata*, Bible), Latvian (novels from http://www.ailab.lv/Teksti/), Malayalam (Bible), Maori (Bible), Russian (Bible), Spanish (novels from www.gutenberg.org), Swedish (Bible), Tagalog (Bible), Turkish (Bible, http://www.cis.hut.fi/morphochallenge2008/datasets.shtml), Vietnamese (Bible)

3.3.1 "Syntactic" languages with few distributional parts of speech

FRENCH

Table 5: Distributional word classes in French (*Les Misérables*)

The 329 (of 30265) wordform types with the most diverse distributions have been classified into 14 distributional parts of speech, all of which are groupings that make sense except for minor outliers (underlined). Labels in left columns are (manual) semantic interpretations a posteriori.

1 Finite verb (3sg)	*allait, appelle, disait, eut, faisait, faut, fit, fut, fût, peut, pouvait, prit, regarda, regardait, répondit, savait, semblait, sentait, serait, soit, tenait, va, venait, vit, voyait*
2 Auxiliary	*a, ai, alla, aurait, avaient, avait, avez, avoir, avons, ayant, eût, font, ont, plein, sont, suis, étaient, étant, êtes, être*
3 Adverb & Past participle	*assez, au-dessus, aussi, autour, dit, déjà, encore, entre, eu, fait, là, passé, pris, pu, toujours, vu, été*
4 Proper names	*ceci, cosette, courfeyrac, enjolras, fantine, fauchelevent, gavroche, gillenormand, javert, jondrette, laquelle, lui-même, m., madeleine, marius, napoléon, paris, thénardier, ça*
5 Subject pronouns	*cela, elle, elles, il, ils, je, lui, on, tu*

6 Prepositions, conjunctions, interrogatives	*ailleurs, au, aux, avec, bien, chez, comme, contre, d', dans, de, des, devant, donc, dont, du, en, est, et, j', jusqu', ou, où, par, pas, plus, pour, qu', que, qui, sans, si, sous, sur, tout, un, valjean*[a)]*, vers, à, était*
7 Attributive NP elements and oblique forms of personal pronouns	*aucun, ce, ces, cet, cette, chaque, jean*[a)]*, l', la, le, les, leur, m', ma, me, mes, mon, n', ne, notre, nous, quel, quelle, quelque, s', sa, se, ses, son, te, toute, toutes, très, une, votre, vous, y*
8 V-initial nouns singular	*air, autre, eau, effet, enfant, esprit, heure, homme, idée, oeil, ombre, âme, évêque*
9 Feminine nouns singular	*barricade, chambre, chose, face, femme, fille, fois, force, lumière, main, maison, mort, mère, nuit, peine, personne, pierre, place, porte, première, rue, terre, tête, vie, ville, voix*
10 Quantifiers and infinitives	*beaucoup, coup, dieu, dire, eux, faire, madame, moi, moins, monsieur, même, parler, passer, peu, point, prendre, quoi, rien, tous, voir*
11 Nouns plural	*ans, autres, bas, bras, cheveux, choses, enfants, femmes, filles, francs, gens, heures, hommes, jours, mains, petits, pieds, yeux*
12 Masculine nouns singular	*bruit, coeur, côté, droit, feu, froid, haut, jardin, jour, lit, mois, moment, monde, mot, mouvement, mur, nom, peuple, premier, père, regard, reste, rire, silence, temps, visage*
13 Numerals	*cent, cinq, deux, grands, leurs, mille, quatre, quelques, six, trois*
14 Adjectives (pre- and postposed!)	*aller, beau, bon, bonne, debout, fort, grand, grande, jeune, mal, mieux, noir, pauvre, petit, petite, possible, près, seul, sombre, trop, venu, vieille, vieux*
15 Adverbs	*ainsi, alors, après, avant, comment, depuis, derrière, enfin, ici, jamais, maintenant, mais, ni, non, pendant, peut-être, pourquoi, pourtant, presque, puis, quand, quelquefois, reprit, seulement, tant, voilà*

[a)] *Jean Valjean* is the only frequent sequence of first and last name in the corpus. This is why these two forms are missclassified.

The applied clustering method (`pam()`,Kaufman & Rousseeuw 2005) is hard clustering. Each form is attributed to one cluster. Thus, homonyms are sorted only into one cluster. However, homonyms can be detected because they change groups with different number of the variable *k* (number of clusters). For instance, in a partitioning with k=20 there is a cluster

18 *ans fois francs heures mille mois sous*

with the distributional part of speech "units of measurement" which has attracted the homonym *sous* 'under; coins'. Consider the forms with a frame in Table 5 where they belong to different parts of speech: plural nouns, feminine nouns, masculine nouns, numerals, and prepositions. As far as distribution is concerned, *mois* and *fois* are equally homonymous as *sous*.

Each emerging cluster is a distributional part of speech. Thus, "units of measurement" are a distributional part of speech in the French corpus, even though it does not emerge with 15 clusters. The appropriate question is not: Are "units of measurement" a part of speech in French, but at which point does this cluster emerge (put differently, which clusters do emerge more easily and which clusters are less prone to appear).

Homonyms are forms which are attributed to different distributional parts of speech in different partitionings. In order to establish homonyms, the two (or three) clusters to which they belong alternatively must be established first.

Thus, Evans & Osada (2005b: 451) are wrong in arguing that semantic heuristics is needed to avoid distributional chaos. There is no distributional chaos, if the right kinds of partitioning algorithms are applied. (Of course, there can be other motivations in favor of semantic heuristics, but the argument of Evans & Osada is wrong).

ENGLISH
Hockett's NV class emerges, but it is a less salient cluster than V(infinitive/present.non3sg) and N. With 14 clusters we get among other things:

6 Noun I [SG/PL]	*air, answer, attention, bed, business, care, coming, conversation, days, fear, feeling, fire, friends, happiness, hope, interest, ladies, life, love, men, money, others, pain, part, pleasure, present, reason, rest, return, sense, silence, smile, thoughts, trouble, turn, walk, wish, women, words, work,*
7 Noun II	*boy, change, child, day, door, evening, gentleman, girl, hour, house, lady, light, man, means, moment, morning, night, person, place, point, question, room, thing, time, town, way, woman, word, world,*
5 Noun III [SG/PL] (inalienable ?): body parts, kin	*countenance, eye, eyes, face, father, friend, glance, hair, hand, hands, head, heart, home, look, manner, master, mind, mother, name, nature, side, step, tone, voice,*
8 Adjectives	*bad, black, certain, close, cold, dark, dead, deep, doing, far, full, glad, good, great, happy, hard, heavy, high, kind, large, late, long, low, mean, mine, new, open, pale, poor, pretty, quiet, real, right, short, silent, small, sound, strange, strong, sure, sweet, talk, true, white,*
13 Verbs	*ask, believe, call, do, feel, find, get, give, hear, help, keep, know, leave, let, make, say, see, show, speak, take, tell, think, want,*

With 15 Cluster NV(A) emerges as a distributional part of speech of its own:

answer, change, fear, hope, look, love, mean, return, talk, turn, walk, want, wish

Does English have a distributional part of speech NV? No, if we want to have less than 15 clusters. Yes, if we want to have 15 clusters or more.
There are many more parts of speech than usually assumed for English. There is, for instance a cluster that comes close to inalienable nouns (nouns which typically occur with possessive pronouns)

Table 6: 489 English wordforms in a literary corpus (J. Austen, Ch. Brontë, Ch. Dickens, M. Twain) classified into 17 distributional parts of speech

Verb infinitive /PRS.NON3SG	*ask, believe, call, do, feel, find, get, give, hear, help, keep, know, leave, let, make, say, see, show, speak, take, tell, think*
Verb past, participle	*answered, asked, brought, called, done, entered, felt, followed, found, gave, given, got, heard, held, kept, knew, known, laid, left, made, opened, passed, put, read, said, saw, seen, set, taken, thought, told, took*
Verb	*appeared, began, came, come, comes, continued, cried, drew, fell, go, lay, looked, looking, ran, remained, replied, returned, rose, run, sat, says, spoke, stood, struck, turned, walked, went*
Conjunctions, Auxiliary, Intensifiers, Prepositions	*all, and, are, as, be, been, but, by, for, had, has, have, how, if, in, is, not, of, or, so, than, that, to, too, very, was, were, what, when, which, with*
Prepositions and verb particles	*about, above, after, against, along, among, at, away, back, before, behind, between, beyond, down, from, having, into, like, near, off, on, out, over, quite, round, seemed, through, together, towards, under, up, upon, without*
Attributes	*another, any, being, each, every, few, five, further, great, hardly, immediately, last, least, less, little, many, miss, more, most, old, other, own, past, same, six, some, such, ten, themselves, these, those, three, tom, two, whole, whose*
Noun I singular	*boy, child, day, door, evening, gentleman, girl, hour, house, lady, man, moment, morning, night, person, place, point, question, room, second, sound, thing, time, way, woman, word, world,*
Noun plural	*days, hours, ladies, men, minutes, people, things, times, women, years*
Noun II (abstract?)	*air, attention, bed, business, care, conversation, feeling, fire, friends, happiness, interest, life, light, means, money, others, pain, pleasure, present, reason, rest, sense, silence, step, town, trouble, words, work*
Noun III (inalienable?)	*countenance, eye, eyes, face, father, friend, glance, hair, hand, hands, head, heart, home, manner, master, mind, mother, name, nature, part, side, smile, thoughts, tone, voice,*
VN	*answer, change, fear, hope, look, love, mean, return, talk, turn, walk, want, wish*
Adjective	*bad, best, black, certain, close, cold, coming, dark, dead, dear, deep, doing, far, full, glad, good,*

	happy, hard, heavy, high, kind, large, late, long, lost, low, new, next, open, pale, particular, poor, pretty, quiet, real, right, short, silent, small, strange, strong, sure, sweet, true, white, young
Auxiliary	*am, can, can't, cannot, could, couldn't, did, didn't, don't, i'll, may, might, must, shall, should, will, would, wouldn't*
Adverb	*afterwards, again, ain't, almost, also, always, became, because, both, does, even, first, half, here, indeed, it's, making, neither, nor, now, often, once, only, perhaps, really, scarcely, seeing, since, sometimes, soon, still, suddenly, taking, then, though, thus, till, where, whether, while, whom, why, within, yet*
Adverb II, reflexive pronoun	*alone, already, anything, better, either, enough, ever, everything, going, gone, herself, himself, just, mine, mr, mrs, much, myself, nobody, none, nothing, oliver, one, rather, sikes, something, there, well, yourself*
Articles, oblique personal pronouns, possessive pronouns	*a, an, her, him, his, its, me, my, no, our, the, their, them, this, us, your*
Personal pronouns, non-oblique	*he, i, it, never, she, they, we, who, you*

TAGALOG

Table 7: Distributional parts of speech in Tagalog (Bible, 303 forms clustered into 28 clusters)

Personal pronoun oblique	*akin* 1SG, *inyo* 2PL, *iyo* 2SG, *kanila* 3PL, *kaniya* 3SG
Personal pronoun oblique + linker *-ng*, Plural word	*aking* 1SG, *aming* 1PL.EX, *ating* 1PL.IN, *inyong* 2PL, *iyong* 2SG, *kanilang* 3PL, *kaniyang* 3SG, *mga* PLURAL.WORD
Personal and demonstrative pronoun non-focus	*ko* 1SG, *mo* 2SG, *namin* 1PL.EX, *natin* 1PL.IN, *nila* 3PL, *ninyo* 2PL, *nito* 'this', *niya* 3SG, *niyaon* 'that'
Personal and demonstrative pronoun focus	*ako* 1SG, *dito* 'this', *doon* 'that', *huwag* 'don't', *ikaw* 2SG, *ka* 2SG.ENCL, *kami* 1PL.EX , *kayo* 2PL, *kita* 1DU/2SG>1SG.NFOC, *sila* 3PL, *siya* 3SG, *tayo* 2DU
Personal and demonstrative pronoun focus/non-focus + linker *-ng*	*akong* 1SG.F, *kang* 2SG.F.ENCL, *kayong* 2PL.F, *kong* 1SG.NF, *mong* 2SG.NF, *nilang* 3PL.NF, *ninyong* 2PL.NF, *niyang* 3SG.NF, *silang* 3PL.F, *yaong* 'that.F'
Numeral + linker	*dalawang* 2, *limang* 5, *pitong* 7, *sangpung* 10, *tatlong* 3
Nouns: proper names, places, times	*egipto*, *gabi* 'night', *ilang* 'desert', *israel*, *jacob*, *jerusalem*, *juda*, *langit* 'sky'
Nouns	*ama* 'father', *asawa* 'husband/wife', *bahay* 'house', *bibig* 'mouth', *ina* 'mother', *kaharian* 'kingdom', *kaluluwa* 'soul', *kamay* 'hand', *lakad* 'walk/manner', *lingkod* 'servant', *mukha* 'face', *pangalan* 'name', *puso* 'heart', *sarili* 'self', *tinig* 'voice', *ulo* 'head'
Nouns and adjectives	*apoy* 'fire', *dagat* 'sea', *ginto* 'money, gold', *iba* 'other', *isa* 'one', *kahoy* 'wood', *kapayapaan* 'peace', *katotohanan* 'truth', *lupa* 'earth', *mabuti* 'good', *masama* 'same', *matuwid* 'just, right', *pilak* 'silver', *una* 'first'
Attributive elements + linker	*anomang* 'any', *bawat* 'each [no linker]', *buong* 'whole', *dakilang* 'great', *ibang* 'other', *isang* 'one', *mabuting* 'good', *malaking* 'strong', *maraming* 'many', *masamang* 'same', *sariling* 'self', *sinomang* 'any', *unang* 'first'
Verb, actor focus (various aspects)	*dumating* 'arrive', *gumawa* 'do', *lumabas* 'exit', *mamatay* 'die', *nagsabi* 'say', *nagsalita* 'talk', *namatay* 'kill', *nangyari* 'happen', *pumasok* 'enter', *tunay* 'true/truly', *wala* 'nothing, not have', *yumaon* '?'
Verb, non-actor focus (various aspects)	*dinala* 'bring', *gagawin* 'do', *ginagawa* 'do', *ginawa* 'do', *ibibigay* 'give', *ibinigay* 'give', *inilagay* 'put', *nakita* 'see', *nalalaman* 'contain?', *narinig* 'hear', *nasumpungan* 'find', *papatayin* 'kill', *pinatay* 'kill', *sinalita* 'talk', *sinugo* 'send', *tinawag* 'call'

According to Himmelmann (1991) and Sasse (1993: 201) Tagalog has six morpho-syntactic slots for content words. However, it is shown here that it is possible to go directly from distribution to a classification of form classes where there are around thirty classes of wordforms in an optimal partitioning.

> "Die kategoriell durchweg vagen Inhaltswörter des Tonganischen etwa werden nicht dadurch zu 'Nomina', daß sie in der Absolutiv- oder Ergativposition auftreten" (Sasse 1993: 199).

Why not? Classes of noun-forms and classes of verb-forms can be established on the basis of distribution. If there is no or little morphology, these can be reinterpreted directly as groups of lexemes. Whether there is fluidity is of secondary importance. What matters is actual distribution, not virtual distribution.

GERMAN

Table 8: Mommsen, *Römische Geschichte* I-V,VIII, 480 forms clustered, 26 clusters

Adjectives and Adverbs
alte, eigene, eigentliche, ganze, griechische, grosse, italische, neue, politische, roemische
aelteren, aeltesten, alten, eigentlichen, ersten, grossen, heutigen, kleinen, letzten, naechsten, neuen, oeffentlichen, rechten, reichen, spaeteren, zweiten
allgemeinen, anderen, eigenen, einzelnen, ganzen, griechischen, hellenischen, italischen, latinischen, politischen, roemischen
alles, allmaehlich, bestimmt, daran, entschieden, frueh, gegenueber, gut, lange, leicht, neu, nie, notwendig, oft, rechtlich, schwer, sicher, voellig, vollstaendig
darauf, frueher, ganz, gar, je, nichts, sehr, viel, weit, wenig
allerdings, also, dafuer, dagegen, darum, dasselbe, dennoch, ebenso, endlich, ferner, freilich, indes, insofern, kaum, natuerlich, seitdem, ueberall, vielmehr, zwar
bald, dabei, dadurch, damit, dann, dort, durchaus, namentlich, nun, sogar, sonst, spaeter, spaeterhin, teils, ueberhaupt, vermutlich, vielleicht, wahrscheinlich, weder, zugleich, zunaechst

(Forms on *-er*, *-es*, *-em* happen not to be represented with adjectives in the 480 forms classified)

3.3.2 “Syntactic” languages with many semantic distributional parts of speech

VIETNAMESE

Table 9: Selected distributional parts of speech (Bible, 718 forms clustered into 100 clusters)

Pronominals	*chúa* ‘lord’, *chúng* ‘PL/group’, *con* ‘child/small’, *họ* ‘tribe’, *mình* ‘body/you’, *ngài* ‘you (deity)’, *ngươi* ‘you (inferiors)’, *người* ‘man/CL’, *nó* ‘s/he (arrogant)’, *ta* ‘I/we’, *tôi* ‘I, slave’, *vua* ‘king’,
Body parts (+)	*bầy* ‘display, flock’, *chơn* ‘limit’, *huyết* ‘blood’, *lòng* ‘innards, heart’, *lưỡi* ‘tongue’, *miệng* ‘mouth’, *môi* ‘lip, *mắt* ‘eye’, *mặt* ‘face’, *ngôi* ‘throne, kingship’, *tay* ‘hand’, *thân* ‘body, *thịt* ‘flesh’, *trại* ‘farm, camp’
Topographic nouns	*biển* ‘sea’, *bờ* ‘edge, bank’, *gát* ‘?’, *mây* ‘cloud’, *núi* ‘mountain’, *rừng* ‘forest, wild’, *sông* ‘river’, *đầu* ‘head, front end’,
Animals	*bò* ‘cow, crawl’, *chim* ‘bird, court’, *chiên* ‘sheep’, *lừa* ‘donkey, deceive’, *ngựa* ‘horse’, *thú* ‘quadruped’, *đực* ‘male (animal)’,
Door words	*cửa* ‘door’, *tường* ‘wall, know well’, *vách* ‘partition, wall’
Quantifiers	*các* ‘all’, *mọi* ‘every’, *một* ‘one’, *nhiều* ‘few, much’, *những* ‘PL’
Numerals	*ba* ‘3’, *bảy* ‘7’, *bốn* ‘4’, *hai* ‘2’, *mười* ’10 (not numerated)’, *năm* ‘5’, *sáu* ‘6’, *tám* ‘8’, *tư* ‘4 (following numeral in ten order)’,
Motion verbs, local verbs	*lên* ‘ascend’, *qua* ‘pass’, *ra* ‘exit’, *tới* ‘come’, *vào* ‘enter’, *xuống* ‘descend’, *đi* ‘go’, *được* ‘get’, *đến* ‘come’, *ở* ‘be.at’,
Qualities (+)	*cao* ‘tall, high’, *do* ‘due to, ashes, spy’, *dài* ‘long’, *ngang* ‘wide’, *rất* ‘very’, *sâu* ‘deep’, *thánh* ‘saint, holy’, *tôn* ‘honor, grandchild, family-’
Verbs (perception and others, transitive)	*biết* ‘know’, *có* ‘have, exist’, *làm* ‘do’, *lấy* ‘take’, *nghe* ‘listen, hear’, *nói* ‘talk, say’, *theo* ‘follow’, *thấy* ‘see, feel’, *tin* ‘believe, news’, *xem* ‘look’
Verbs of saying	*hỏi* ‘ask’, *phán* ‘order’, *thưa* ‘reply, thin’, *tiếp* ‘join, continue’, *tưởng* ‘believe, praise, think’, *đáp* ‘answer’
Verbs (transitive)	*bán* ‘sell, half’, *chuộc* ‘buy back’, *cưới* ‘marry’, *giúp* ‘help’, *nộp* ‘deliver’, *quên* ‘forget’, *đóng* ‘close, build’
Transitive auxiliary verbs	*cho* ‘give’, *cùng* ‘accompany’, *với* ‘join’
Negators, tense, modality, focus particles	*bèn* ‘then, instantly’, *chẳng* ‘not, no’, *cũng* ‘also’, *không* ‘not’, *lại* ‘again’, *phải* ‘must, right’, *sẽ* ‘FUTURE’, *đã* ‘PERFECT, already’, *đặng* ‘can, able’, *đều* ‘equal, even’
Prepositions (+)	*cả* ‘all’, *của* ‘POSS’, *dưới* ‘below’, *giữa* ‘middle’, *khỏi* ‘avoid’, *là* ‘equal’, *nơi* ‘place’, *trong* ‘inside’, *trên* ‘above’, *trước* ‘before, front’, *tại* ‘be at’, *từ* ‘from’, *về* ‘about, return’
Conjunctions	*mà* ‘but’, *như* ‘like’, *nên* ‘thus’, *rằng* ‘saying’, *rồi* ‘finish, then, already’, *thì* ‘time, then’, *và* ‘and’, *vì* ‘because, throne’, *để* ‘in order to, put’
Syllables occurring in foreign proper names	*ga, ghê, giê, lê, phê, sa, sê, xê, xô, ê*

INDONESIAN

Table 10: Intermediate between syntactic and semantic, 28 distributional parts of speech (whereof 11 “nouns” and 8 “verbs”)

Mass nouns (incl. animals)	*air* ‘water’, *anggur* ‘wine’, *api* ‘fire’, *batu* ‘stone’, *burung* ‘bird’, *darah* ‘blood’, *domba* ‘sheep’, *dosa* ‘sin’, *emas* ‘gold’, *gandum* ‘wheat’, *hujan* ‘rain’, *kayu* ‘wood’, *kejahatan* ‘crime’, *kuasa* ‘power’, *kurban* ‘sacrifice’, *makanan* ‘food’, *pakaian* ‘cloth’, *perak* ‘silver’, *persembahan* ‘gift’, *sapi* ‘cow’, *uang* ‘money’
Places	*bukit* ‘hill’, *gerbang* ‘?’, *gunung* ‘mountain’, *kemah* ‘tent’, *laut* ‘sea’, *lembah* ‘valley’, *mezbah* ‘?’, *pintu* ‘door’, *pohon* ‘tree’, *tembok* ‘wall’
Place names	*babel, bumi* ‘earth’, *dunia* ‘world’, *filistin, israel, langit* ‘sky’, *lewi, mesir, moab, yahudi, yehuda, yerusalem*
Human nouns	*anaknya* ‘child.3POSS’, *binatang* ‘animal’, *hamba* ‘slave’, *imam* ‘priest’, *istri* ‘wife’, *laki-laki* ‘man’, *malaikat* ‘angel’, *nabi* ‘prophet’, *pemimpin* ‘leader’, *penguasa* ‘ruler’, *perempuan* ‘woman’, *wanita*
Names (persons)	*abraham, ayahnya, daud, harun, musa, paulus, petrus, rakyat, salomo, saul, yakub, yesus, yosua, yusuf*
Classifiers for humans	*anak* ‘child’, *bangsa* ‘people’, *orang* ‘man’, *orang-orang* ‘man:PL’, *penduduk* ‘inhabitant’, *raja* ‘king’, *semua* ‘all’, *seorang* ‘one:man’, *tentara* ‘army’, *umat* ‘COLL/group’
Titles	*baginda* ‘majesty’, *bapak* ‘father’, *dirinya* ‘self:3POSS’, *kau* 2SG, *kristus, manusia* ‘human’, *saudara* ‘brother’, *tuan* ‘lord’, *tuanku* ‘lord:2SG’, *umat-nya* ‘people:3PL’
Personal pronouns (including ‘God’)	*aku* 1SG, *allah* ‘God’, *dia* 3SG, *engkau* 2SG, *ia* 3SG, *kalian* 2PL, *kami* 1PL.EXCL, *kamu* 2.FAM, *kita* 1PL.INCL, *mereka* 3PL, *saya* 1SG, *tuhan* ‘God’
Adjectives	*baik* ‘good, beautiful’, *banyak* ‘much’, *baru* ‘new’, *berani* ‘brave’, *besar* ‘big’, *dekat* ‘near’, *gembira* ‘happy’, *jahat* ‘bad’, *jauh* ‘far’, *kuat* ‘strong’, *lain* ‘other’, *lainnya* ‘other:3POSS’, *sama* ‘same’, *suci* ‘clean, holy’
Numerals, quantifiers	*beberapa* ‘some’, *dua* ‘2’, *empat* ‘4’, *kedua* ‘2nd, both’, *lima* ‘5’, *satu* ‘1’, *setiap* ‘every’, *tiga* ‘3’, *tujuh* ‘7’
Transitive verbs	*melawan* ‘resist, against’, *melihat* ‘see’, *memanggil* ‘call’, *membawa* ‘carry’, *memberkati* ‘bless’, *membiarkan* ‘allow’, *membuat* ‘make’, *membunuh* ‘kill’, *memilih* ‘choose’, *memperhatikan* ‘heed’, *mencari* ‘search’, *mendengar* ‘hear’, *mendengarkan* ‘listen’, *mengalahkan* ‘surpass’, *mengangkat* ‘pick up’, *mengasihi* ‘love’, *mengenal* ‘know’, *menghukum* ‘punish’, *mengikuti* ‘follow’, *mengumpulkan* ‘call together’, *mengutus* ‘send’, *meninggalkan* ‘leave’, *menolong* ‘help’, *menyelamatkan* ‘save’, *menyembah* ‘do hommage’, *menyerahkan* ‘deliver’, *menyerang* ‘attack’, *menyuruh* ‘order’
Motion verbs	*berangkat* ‘depart’, *datang* ‘come’, *jatuh* ‘fall’, *keluar* ‘exit’, *kembali* ‘return’, *lari* ‘run’, *masuk* ‘enter’, *pergi* ‘go’, *pulang* ‘come home’, *turun* ‘descend’
Locative verbs	*ada* ‘exist’, *bekerja* ‘work’, *berdiri* ‘stand’, *berjalan* ‘walk’, *duduk* ‘sit, live’, *hidup* ‘live’, *ikut* ‘take part, together’, *makan* ‘eat’, *memerintah* ‘rule’, *penuh* ‘full’, *sampai* ‘until, arrive’, *tinggal* ‘stay, live’

3.3.3 “Syntactic” languages with many formal distributional parts of speech

ITALIAN

Some parts of speech emerge very well with lower number of clusters and are too subdivided here (e.g., nouns M SG), some few do not even clearly cluster with 53 (infinitives, proper names)

Table 11: 524 forms clustered into 53 parts of speech

Noun M SG	*braccio, capo, corpo, figlio, fine, letto, luogo, mezzo, nome, passo, potere, volto*
Noun M SG	*cielo, conte, garibaldi, giovane, giovine, mondo, nemico, padre, papa, popolo, prete, re, signore, sole*
Noun M SG	*bisogno, cavaliere, cuore, danno, diritto, dolore, figliuolo, fratello, fuoco, male, pensiero, pericolo, sangue*
Noun M SG	*caso, dì, giorno, modo, momento, punto, tempo, tratto*
Noun M PL	*capelli, fatti, figli, giorni, morti, nemici, passi, pensieri, piedi, popoli, soldati, tempi*
Noun F SG	*bocca, casa, chiesa, donna, faccia, francia, fronte, gente, guerra, madre, mano, mente, morte, notte, pace, persona, terra, testa, via, vita, voce*
Noun F SG	*città, famiglia, parola, parte, porta, prova, stanza, storia, volta*
Noun F SG abstract	*fama, fede, fortuna, forza, libertà, luce, natura, paura, pietà, ragione, virtù*
Noun F PL	*altre, armi, braccia, cose, donne, mani, ore, parole*
#V noun/adj SG	*acqua, altra, altro, amico, amore, anima, animo, antica, antico, aria, atto, italia, opera, ultimo, uomo*
#V/sC M Pl	*altri, amici, anni, occhi, stati, uomini*
Adj M SG	*bene, buono, là, morto, nuovo, proprio, subito, vero*
Adj M SG	*cardinale, cènci, pari, posto, primo, quale, santo, secondo*

Adj F SG	*bella, buona, certa, grande, lunga, nuova, povera, santa, sola, stessa*
Adj M SG #V	*bel, buon, gran, mal, povero, signor, vecchio*
Poss F SG	*mia, nostra, propria, sua, tua, vostra*
Poss M SG	*medesimo, mio, nostro, suo, tuo, vostro*
Poss M PL	*miei, nostri, suoi, vostri*
Proper name +	*beatrice, costui, cresti, cristo, damiano, dio, ezio, flora, manfredi, marzio, nessuno, quegli, rogiero*
Proper name PL	*francesi, quali, romani*
Pronoun	*appena, egli, ei, ella, essa, essi, io, tu*
Pron obl emph	*cui, lei, loro, lui, me, noi, roma, sè, te, voi*
Pron obl	*ci, gli, le, li, lo, mi, ne, non, si, ti, vi*
Adv	*almeno, anzi, bensì, certo, chè, ecco, finchè, imperciocchè, intanto, mentre, oggi, onde, poichè, quantunque, tuttavia*
Adv	*addosso, davanti, fino, innanzi, intorno, lì, pertanto*
Adv (mann +)	*breve, forte, giù, grave, insieme, oltre, piuttosto, presso, presto, quivi, spesso, tardi, volte*
Adv (quant)	*assai, ben, cosa, molto, poco, quanto, quasi, questa, questo, sempre, sì, tanto, troppo, tutti, tutto*
Adv (temp +)	*adesso, allora, anco, chi, dunque, forse, già, ormai, però, poi, pure, questi, qui, solo, veramente*
Adv (loc)	*contro, dentro, dietro, dopo, fra, fuori, sopra, sotto, su, tra, tutta, verso*
Quantifiers	*alcuni, altrui, coteste, don, due, mille, molte, pochi, quattro, queste, tanti, tre*
Infinitive +	*credere, dare, dire, far, fare, lungo, mettere, morire, sapere, tutte, vedere*
Infinitive	*andare, aver, avere, essere*
‘be’	*era, erano, fosse, fu, mai, sarebbe, sei, sia, sono, è*
‘have’	*abbia, abbiamo, avesse, avessero, avete, aveva, avevano, avrebbe, avrebbero, ha, hai, hanno, ho*
Verb	*ebbe, fa, faceva, fanno, fece, fossero, sentiva, siete, teneva*
Verb	*appunto, diceva, disse, pareva, parve, rispose, sarà, vide*
Verb	*andava, andò, avendo, furono, pare, prese, rimase, siamo, sta, stava, va, venne, viene*
Modal	*bisogna, deve, doveva, possa, posso, potesse, poteva, potè, può, sa, sapeva, so, son, voleva, volle, vuole*
?	*cotesti, farsi, que', quei*
Partic Aux	*avuto, potuto, voluto*
Partic	*dato, detto, fatta, fatto, maggiore, messo, preso, venuto*
Prep	*a, ad, che, con, da, di, e, in, per, più, senza*
Prep var.	*a', ai, co', dai, dallo, de', dei, esser, fin, nei, nello, pei, pel*
(Prep +)Art M	*al, col, dal, del, i, il, nel, sul, un*
(Prep +) Art F SG	*alla, colla, dalla, della, la, nella, sulla, una*
(Prep +) Art M PL	*agli, dagli, degli, gl', negli*
Prep + F PL	*alle, allo, dalle, delle, nelle, sue*
Prep +Art #V	*all', coll', d', dall', dell', dello, l', nell', quell', un'*
Conj, Adv	*anche, come, così, dove, ma, nè, o, ora, perchè, quando, se*
Conj /_#V	*anch', ch', ed, m', s'*
?	*federigo, francesco, giovanni, lontano, molti, peggio, vivere*
?	*carlo, ciò, meglio, meno, nulla, prima, quella, quelle, quelli, quello, tale*
?	*cotesta, cotesto, ogni, qual, qualche, quel, san, tanta, uno*
? #V/sC	*ancora, imperatore, pur, spirito, stata, stato, stesso*

LATIN

Table 12: Some examples for 88 clusters of the 588 forms in the Vulgata with the most different distributions (best number of clusters)

Noun GEN M SG (animate)	*christi* ‘Christ’, *dei* ‘God’, *domini* ‘lord’, *hominis* ‘man’, *ipsius* ‘self’, *patris* ‘father’, *regis* ‘king’
Noun ACC F SG (inanimate)	*animam* ‘soul’, *domum* ‘house’, *faciem* ‘appearance’, *manum* ‘hand’, *partem* ‘part’, *viam* ‘way’, *vocem* ‘voice’
Noun ABL (SG/PL) (M/N/F)	*conspectu* ‘appearance’, *corde* ‘heart’, *manibus* ‘hand.PL’, *manu* ‘hand’, *medio* ‘middle’, *nomine* ‘name’, *oculis* ‘eye.PL’, *ore* ‘mouth’, *parte* ‘part’, *peccato* ‘sin’, *sanguine* ‘blood’
Proper names	*aaron*, *beniamin*, *ephraim*, *manasse*, *moab*, *philisthim*
Proper names II	*abraham*, *absalom*, *altare* ‘altar.ABL’, *david*, *iacob*, *ioab*, *ionathan*, *ioseph*, *iosue*, *moyses*, *samuel*, *saul*
Possessive pronouns, DAT/ABL.PL	*meis* 1SG, *nostris* 1PL, *suis* 3, *tuis* 2SG, *vestris* 1PL
Personal/demonstrative pronouns DAT	*ei* ‘him/her’, *eis* ‘them’, *illi* ‘that.SG’, *illis* ‘that.PL’, *mihi* 1SG, *tibi* 2SG
Numerals except tens	*centum* 100, *duo* 2, *duodecim* 12, *duos* 2:ACC.M, *quattuor* 4, *quinque* 5, *septem* 7,

	sex 6, *tres* 3, *tribus* 3.OBL
Numerals tens	*decem* 10, *quadraginta* 40, *quinquaginta* 50, *triginta* 30, *viginti* 20
Prepositions with accusative	*ad* ‘to’, *circa* ‘around’, *contra* ‘against’, *forte* ‘strong[ADV]’, *per* ‘for’, *super* ‘over’
Prepositions with ablative (and accusative)	*a* ‘from’, *ab* ‘from’, *cum* ‘with’, *de* ‘down from’, *ex* ‘out of’, *hora* ‘hour[NOM/ABL]’, *in* ‘in’, *patrum* ‘father.GEN.PL’, *prae* ‘in front’, *pro* ‘for’, *sine* ‘without’, *sub* ‘under’
Prepositions with accusative	*iuxta* ‘along’, *omnem* ‘all.ACC.SG’, *propter* ‘because of’, *secundum* ‘following’

RUSSIAN

In a partitioning with 117 distributional parts of speech, 13 contain possessive pronouns (rows in Table13). Each case/gender/number form of possessive pronouns is a distributional part of speech of its own. Russian has 3 genders 2 numbers and 6 (major) cases. If there would not be any syncretism, the part of speech system would exhibit a much higher complexity (2 * 3 * 6 = 36 form-classes). Syncretism has the effect of making the part of speech system considerably less complex. (Complexity: number of distributional parts of speech required). Attributive adjectives are grouped with possessive pronouns. The forms included here testify to the particular nature of the corpus (Bible). It is not clear whether a corpus of colloquial Russian with different distributional properties of possessive pronouns would yield the same result. Table 13 shows how lexemes and word classes can be extracted from distributional parts of speech. The rows are obtained distributional parts of speech. These can be easily grouped to a paradigm as done here. The columns are the lexemes, the whole table is a word class and a paradigm. In order to perform the lexeme analysis, some morphological analysis is required. It is, however, much easier to align a set of distributional parts of speech into a set of lexemes than to find lexemes in the set of all forms. There is a single misclassified element here: *вся*.

Table 13: Selected distributional parts of speech in Russian (rows), paradigmatized manually (columns)

Interpretation	‘God:ADJ’	POSS2PL	‘all’	‘Lord:ADJ’	‘Israel.ADJ’	POSS1SG	POSS1PL	POSS3RFL	‘this’	POSS2SG
NOM(ACC) SG M		ваш			израилев	мой	наш	свой		твой
NOM/ACC SG N	божие	ваше		господне		мое		свое	сие	твое
NOM SG F						моя				твоя
GEN/ACC SG M/N	божия	вашего		господня	израилева	моего	нашего	своего		твоего
GEN/DAT SG F						моей		своей		твоей
DAT SG M/N				господню		моему		своему	сему	твоему
ACC SG F	божию					мою		свою	сию	твою
INS SG M/N, DAT.PL		вашим			израилевым	моим	нашим	своим		твоим
INS SG F			~~вся~~					своею		твоею
LOC SG M/N						моем		своем		твоем
NOM(ACC) PL		ваши			израилевы	мои	наши	свои		твои
GEN(ACC)/LOC PL		ваших			израилевых	моих	наших	своих		твоих
INS PL			всеми			моими		своими		твоими

Forms of quantifiers and demonstrative pronouns also occur in other clusters not given here. All 117 clusters make sense to a certain extent (but many contain some misclassifications). Table 14 lists some of the verb form clusters obtained:

Table 14: Russian verbal distributional parts of speech (selected)

Past M SG (underlined PRS3SG)	взял, возвратился, вошел, встал, вышел, идет, пойдет, пошел, придет, пришел, сделал, стал, умер
Infinitive	взять, говорить, делать, жить, идти, пить, служить, ходить
Converb	взяв, видя, оставили, сделаю, увидев, услышав
Past PL	взяли, видели, делали, ели, сделали, увидели

The result violates Nau’s P.6 “Few classes” criterion. However, distributional parts of speech that make sense only occur with a high number of k in languages such as Russian and Latin. Morphology is needed to paradigmatize the distributional parts of speech into a smaller number of word classes.

3.3.4 Distributional parts of speech in morphological languages

TURKISH

Partitioning Turkish wordforms according to distributional criteria does not yield a set of clusters which all make sense. However, some clusters are good, for instance, numerals and case forms of personal pronouns. Case forms do not emerge as distributional parts of speech with nouns, but only with personal and demonstrative pronouns

Table 15: Turkish distributional parts of speech (Bible, 68 clusters, 1006 forms clustered)

Numerals	*altı* 6, *aynı* ‘the same’, *beş* 5, dört 4, *geçen* ‘past’, *iki* 2, *sekiz* 8, *yedi* 7, *yetmiş* 70, *üç* 3
Numerals (tens)	*bin* 1000, *elli* 50/‘with hands’, *kırk* 40, *on* 10, *otuz* 30, *yirmi* 20, *yüz* 100, *öbür* ‘other (of two)’
Ordinal Numerals	*birinci* ‘1st’, *ikinci* ‘2nd’, *ilk* ‘primary’, *yedinci* ‘7th’, *üçüncü* ‘3rd’
Mass nouns	*ay* ‘month/moon’, *da* ‘also/but’, *gece* ‘night’, *gün* ‘day’, *gün-ler* ‘day-PL’, *gün-ler-de* ‘day-PL-LOC’, *gün-ü* ‘day.POSS3’, *hal-de* ‘state-LOC’, *hiç* ‘absolutely, nothing’, *kent-te* ‘city-LOC’, *kişi* ‘person’, *kişi-yi* ‘person-POSS3’, *yıl* ‘year’, *zaman* ‘time’, *şey-i* ‘thing-POSS3’
Pronouns ACC	*beni* ‘1SG.ACC’, *biri* ‘one.ACC’, *bizi* ‘1PL.ACC’, *bun-lar-ı* ‘this-PL-ACC’, *bunu* ‘this.ACC’, *bütün* ‘whole’, *insan* ‘human.being’, *israilli-ler-in* ‘Israelit-PL-GEN’, *kendi-ni* ‘self.ACC’, *on-lar-ı* ‘that-PL-ACC’, *onu* ‘that.ACC’, *seni* ‘2SG.ACC’, *sizi* ‘2PL.ACC’, *yine* ‘again’
Pronouns COMIT	*benimle* ‘with 1SG’, *bilge* ‘?’, *bizimle* ‘with 1PL’, *davuta* ‘David-DAT’, *el-in-de* ‘hand-POSS3-LOC’, *hep* ‘all, whole’, *hepsi-ni* ‘all.together-ACC’, *israili*, *musayla* ‘with Moses’, *nereye* ‘whither’, *onlarla* ‘with those’, *onunla* ‘with that’, *seninle* ‘with 2SG’, *sensin* ‘it is you’, *sizinle* ‘with 2PL’, *soyunu* ‘family.POSS3.ACC’, *yanına* ‘to’
Pronouns GEN	*benim* ‘1SG.GEN’, *bizim* ‘1PL.GEN’, *bunun* ‘this.GEN’, *gücü* ‘force-POSS3’, *kendi-si* ‘self.POSS3’, *para* ‘money’, *savaş* ‘war, fight’, *sizin* ‘1PL.GEN’, *sonsuz* ‘infinite’, *yap-an* ‘make-PTC.PRS’, *yol-da* ‘way-LOC’, *zeytinyağı* ‘olive-ADJ’
Pronouns NOM	*ama* ‘but’, *ben* ‘1SG’, *bu* ‘this’, *ey* ‘hey!’, *isa* ‘Jesus’, *kral* ‘king’, *neden* ‘why?’, *o* ‘that’, *orada* ‘there’, *rab* ‘lord’, *sonra* ‘after’, *tanrı* ‘God’, *çünkü* ‘because’,
Pronouns NOM	*baba* ‘father’, *biz* ‘1PL’, *efendim* ‘lord.POSS1SG’, *hem* ‘even, but’, *herkes* ‘every’, *kim* ‘who’, *kimse* ‘somebody’, *mesih* ‘Messiah’, *on-lar* ‘3-PL, *sen* ‘2SG’, *siz* ‘2PL’
Pronouns ABL	*az* ‘little’, *benden* ‘1SG.ABL’, *biraz* ‘little’, *bizden* ‘1PL.ABL’, *bundan* ‘this.ABL’, *bunlardan* ‘these.ABL’, *bunların* ‘those.GEN’, *burada* ‘here’, *buraya* ‘hither’, *gök-ten* ‘sky-ABL’, *haydi* ‘come!’, *israil-de* ‘Israel-LOC’, *kendi-sin-den* ‘self-POSS3-ABL’, *kötü-ler* ‘bad-PL’, *mısır-da* ‘Egypt-LOC’, *ondan* ‘that.ABL’, *onlardan* ‘those.ABL’, *rab-den* ‘lord-ABL’, *senden* ‘2SG.ABL’, *sizden* ‘2SG.ABL’, *toplam* ‘sum’, *topluluğ-u* ‘whole group-POSS3’, *yanında* ‘at’, *ölüm-den* ‘death-ABL’

Verb forms do not cluster according to form-classes, but rather according to lexemes, this does not astonish with the auxiliary *et-*, since auxiliaries and modals also cluster in predominantly syntactic languages.

edecek, *edeceğim*, *eder*, *edin*, *ediyor*, *et*, *etti*, *ettiler*

Other verbs, however, also cluster according to lexeme or semantic group: *yap-* ‘make’

yap, *yapacak*, *yapacağım*, *yaptı*, *yaptılar*, *yapın*

Motion verbs: *dön-* ‘return’, *gel-* ‘come’, *git-* ‘go’, *gönder-* ‘send’, *götür-* ‘bring away’, *çak-* ‘exit’

döndü, *geldi*, *geldiler*, *gidecek*, *git*, *gitti*, *gittiler*, *gönderdi*, *götürdü*, *çağırdı*

-> Pronouns tend to have flectional behavior even in languages where the noun is agglutinative. This correlates with the fact that pronominal forms can be extracted as distributional parts of speech.

Similarly, Finnish, Hungarian, and Malayalam

4. Conclusions

- Functional domain-based typology has been very successful in completely avoiding parts of speech. Croft has reinterpeted such an avoiding approach as a theory of parts of speech. Most non-avoiding approaches to parts of speech by typologists are not efficient (very few languages classified). The only large-scale approach (Hengeveld, Rijkhoff) is not generally accepted. There is hardly any basic field in typology where there is more disagreement than parts of speech. This is the only general conclusion that can be drawn.
- Parts of speech are language-specific (against Croft). Universals of part of speech are procedural rather than structural. Classes should be obtained (bottom-top) rather than given (Nau 2001: 26).
- All languages must have some parts of speech (more than one, minimum hardly below four or five). This follows from the LNRE nature of language (3.1.4). This conclusion is not compatible with the approaches by Hengeveld, by Sasse, and by Gil. For the same reason, parts of speech must be units of language, not only of linguistics. This conclusion is not compatible with the approach by Nau.
- Parts of speech of some kind are a necessary stage in acquisition. This stage precedes syntax; general construction patterns cannot be built without some partitioning of wordforms into groups. In fact, virtually all theories of syntax take some word classes for given.
- Classifying wordforms (or lexemes) into parts of speech (or word classes) is partitioning. Partitioning is cluster analysis which is a field of statistics.
- All approaches to word classes discussed here except Harris and Biemann (and to a certain extent Crystal) underestimate the potential of distribution. This is due to a lack of understanding of cluster analysis (partitioning as a technical term). Statements by Bloomfield, Croft, Evans & Osada and many others about the limits of a purely distributional approach are completely mistaken because they do not take into account cluster analysis.
- Distributional parts of speech (classes of wordforms obtained by distributional analysis from a corpus) can account for word classes in a large number of languages with little morphology, but also in at least some languages with mainly flective morphology (Latin, Russian).
- Distributional parts of speech can also account for some aspects of word class systems in morphological languages, such as Turkish. It has been illustrated in 3.3.4 that they are applicable in those areas of morphological languages where there is little or no overt morphological marking.
- Corpora of at least 1 m words can be classified according to whether their most frequent wordforms can all be partitioned into parts of speech that make sense on the basis of distribution alone. Work in progress suggests that there is a typological difference between syntactic languages and morphological languages, which are defined as follows:

SYNTACTIC LANGUAGES	MORPHOLOGICAL LANGUAGES
Word classes and lexemes can be obtained from distributional parts of speech which can be derived from distribution in a large corpus	Word classes and lexemes cannot be derived (or only to a certain extent) without some morphological analysis
Haitian Creole, English, French, Swedish, Tagalog, Maori, Indonesian, Italian, Spanish, German, Bulgarian, Albanian, Greek, Latin, Russian, Vietnamese	Turkish, Hungarian, Finnish, Estonian, Malayalam, West Greenlandic

  ? Latvian (distributional parts of speech do not work very well even though Latvian has not more morphology than Latin and Russian)

  (There is no strict distinction. Finnish is more syntactic than Hungarian and Turkish. In some languages with syntactic distributional parts of speech, such as Italian, Russian, and Latin, morphology is needed to cluster parts of speech into word classes).
  This typology is efficient in the sense that any language where there are several corpora of ~1 m words freely available can be tested easily. However, it is not efficient in practice since for most languages such corpora are not available.

- WHAT IS NEEDED TO SURVEY PARTS OF SPEECH IS FIRST OF ALL LARGE (ELECTRONIC) CORPORA. Reference grammars are completely useless for determining distributional parts of speech since the actual distribution of wordforms is not represented in grammar.

It is represented in text. Languages replicate from the application of a universal algorithm to language-particular input which consists of a mass of utterances (a corpus). Languages cannot replicate from a description or a set of rules.

- Word classes in morphological languages cannot be studied without some previous analysis of morphology. But the same approach that is applied to wordforms in syntactic languages can be applied to morpheme classes once morphemes have been segmented.
- In order to study word classes in morphological languages we must study morpheme classes which can be done only if we have corpora with morpheme segmentation (whether segmented manually or by means of unsupervised learning approaches).
- To my astonishment, distributional parts of speech are very much like what tradition suggests word classes to be like, not supporting the criticism of Steblin-Kaminsky's bonmot quoted in 1.1. Particularly robust in most languages considered are numerals and pronouns. (Frequently encountered are also proper names, auxiliaries, and modal verbs.) This is some evidence in favor of Haspelmath's saying: *die meisten Linguisten machen alles ungefähr richtig* (p.c.)

References

Ansaldo, Umberto & Don, Jan & Pfau, Roland (eds.). 2008. Parts of speech: Descriptive tools, theoretical constructs. Special Issue of *Studies in Language* 32.3: 505-785.

Anward, Jan & Moravcsik, Edith & Stassen, Leon. 1997. Parts of speech: A challenge for typology. *Linguistic Typology* 1: 167-183.

Baayen, R. H. 2001. *Word Frequency Distributions*. Dordrecht: Kluwer.

Baayen, R. H. 2008. *Analyzing Linguistic Data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bloomfield, Leonard. 1933. *Language*. New York: Holt.

Biemann, Chris. 2006a. Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. *Proceedings of the Workshop on Textgraphs at the HLT/NAACL*, New York City, NY, USA, June.

Biemann, Chris. 2006b. Unsupervised part-of-speech tagging employing efficient graph clustering. *Proceedings of the Student Research Workshop at the COLING/ACL*, Sydney, Australia, July.

Bordag, Stefan. 2007. Elements of knowledge-free and unsupervised lexical acquisition. Dissertation. Fakultät für Mathematik und Informatik der Universität Leipzig. Leipzig.

Bowern, Claire. 2008. *Linguistic Fieldwork: a practical guide*. Basingstoke: Palgrave Macmillan.

Broschart, Jürgen. 1997. Why Tongan does it differently: Categorial distinctions in a language without nouns and verbs. *Linguistic Typology* 1: 123-165.

Chitashvili, R. J. & Khmaladze, E. V. Statistical analysis of large number of rare events and related problems. *Transactions of the Tbilisi Mathematical Institute* 91: 196-245.

Croft, William. 2000. *Explaining Language Change. An evolutionary approach.* Harlow: Longman.

Croft, William. 2001. *Radical Construction Grammar: Syntactic theory in typological perspective.* Oxford: Oxford University Press.

Croft, William. 2005. Word classes, parts of speech, and syntactic argumentation. *Linguistic Typology* 9.3: 431-441.

Crystal, David. 1967. Word classes in English. *Lingua* 17: 24-56.

Cysouw, Michael & Biemann, Chris & Ongyerth, Matthias. 2007. Using Strong's Numbers in the Bible to test an automatic alignment of parallel texts. *Sprachtypologie und Universalienforschung STUF* 60.2: 158-171.

Dennett, Daniel Clement. 1995. *Darwin's Dangerous Ideas. Evolution and the meanings of life*. London: Penguin.

Dik, Simon C. 1989. *The Theory of Functional Grammar*. Part I: *The structure of the clause*. (Functional Grammar Series 9.) Dordrecht: Foris.

Dixon, Robert M. W. 1977. Where have all the adjectives gone? *Studies in Language* 1: 19-80.

Dixon, Robert M. W. 2004. Adjective classes in typological perspective. In Dixon, R. M. W. & Aikhenvald, Aleksandra Y. (eds.), *Adjective Classes. A cross-linguistic typology*. Oxford: Oxford University Press.

Dryer, Matthew S. 2005. Order of adjective and noun. WALS Chapter 89.

Evans, Nicholas. 2000. Word classes in the world's languages. In Booij, Geert & Lehmann, Christian & Mugdan, Joachim (eds.) Morphologie / Morphology. Ein internationales Handbuch zur Flexion und Wortbildung I: 708-732. Berlin: de Gruyter.

Evans, Nicholas & Osada, Toshiki. 2005. Mundari: The myth of a language without word classes. *Linguistic Typology* 9.3: 351-390.

Evans, Nicholas & Osada, Toshiki. 2005b. Mundari and argumentation in word-class analysis. *Linguistic Typology* 9.3: 442-457.

Fries, Charles Carpenter. 1952. *The Structure of English. An introduction to the construction of English sentences*. New York: Harcourt.

Garde, Paul. 1981. Des parties du discours, notamment en russe. *Bulletin de la Société de Linguistique de Paris* 76: 155-189.

Gil, David. 2000. Syntactic categories, cross-linguistic variation and universal grammar. In P. Vogel & B. Comrie (eds), 173-216.

Gil, David. 2005. Adjectives without nouns, WALS chapter 61.

Gil, David. 2008. The acquisition of syntactic categories in Jakarta Indonesian. *Studies in Language* 32.2: 637-669.

Givón, Talmy. 1979. *On Understanding Grammar*. New York: Academic Press.

Givón, Talmy. 1981 Typology and functional domains. *Studies in Language* 5: 163-193.

Harrell, Frank E. 2001. *Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis.* New York: Springer.

Harris, Zellig Sabbettai. 1946 / 1981. From morpheme to utterance. *Language* 22.3: 161-183 / In Harris, Z. S., *Papers on Syntax*, 45-70. Dordrecht: Reidel.

Harris, Zellig Sabbettai. 1951. *Methods in Structural Linguistics*. Chicago: University of Chicago Press. / Reprinted 1960 as *Structural Linguistics*, Phoenix Books.

Hengeveld, Kees. 1992. *Non-Verbal Predication. Theory, typology, diachrony*. Berlin: Mouton de Gruyter.

Hengeveld, Kees & Rijkhoff, Jan. 2005. Mundari as a flexible language. *Linguistic Typology* 9.3: 406-431.

Himmelmann, Nikolaus. 1991. *The Philippine challenge to Universal Grammar*. Arbeitspapiere des Instituts für Sprachwissenschaft der Universität zu Köln, Neue Folge 15.

Hockett, Charles F. 1958. *A Course in Modern Linguistics*. New York: Macmillan.

Hopper, Paul J. & Thompson, Sandra A. 1984. The discourse basis for lexical categories in universal grammar. *Language* 60: 703-752.

Kaufman, Leonard & Rousseeuw, Peter J. 2005. *Finding Groups in Data. An introduction to cluster analysis.* Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley.

Lehmann, Christian. 1982. Directions for interlinear morphemic translations. *Folia Linguistica* 16: 199-224.

Lichtenberk, Frantisek. 1991. Semantic change and heterosemy in grammaticalization. *Language* 67: 474-509.

Lindblom, Björn. 1992. Phonological units as adaptive emergents of lexical development. In Ferguson, C. A. & Menn, L. & Stoel-Gammon, C. (eds.) *Phonological Development: models, research, implications*, 131-163. Timonium, MD: York Press.

Mauthner, Fritz. 1923 / 1982. *Beiträge zu einer Kritik der Sprache*. Erster Band: *Zur Sprache und zur Psychologie*. 2., vermehrte Auflage. Leipzig: Meiner / Frankfurt: Ullstein Materialien.

Mayer, Thomas & Wälchli, Bernhard. 2007. A typological approach to algorithmic morphology. Paper hold at the Algomorph Workshop, Konstanz 3/4.11.2007. http://typo.uni-konstanz.de/algomorph/inc/intro.pdf

Miestamo, Matti. 2005. *Standard Negation: The negation of declarative verbal main clauses in a typological perspective.* (Empirical Approaches to Language Typology 31.) Berlin: Mouton de Gruyter.

Miestamo, Matti. 2007. Symmetric and asymmetric encoding of functional domains, with remarks on typological markedness. In Miestamo, Matti & Wälchli Bernhard (eds.), *New Challenges in Typology: Broadening the Horizons and Redefining the Foundations*, 293-314. Berlin: de Gruyter.

Mosel, Ulrike & Hovdhaugen, Even. 1992. *Samoan Reference Grammar*. Oslo: Scandinavian University Press, The Institute for Comparative Research in Human Culture.

Nau, Nicole. 2001. Wortarten und Pronomina, Studien zur lettischen Grammatik. Habilitationsschrift. Kiel: Universität Kiel.

Peterson, John. 2005. There's a grain of truth in every "myth", or, Why the discussion of lexical classes in Mundari isn't quite over yet. *Linguistic Typology* 9.3: 391-405.

Plank, Frans. 1997. Word classes in typology. Recommended reading. Linguistic typology 1: 185-192.

Rijkhoff, Jan. 2002. *The Noun Phrase*. Oxford: Oxford University Press.

Sapir, Edward. 1921. *Language*. New York: Harcourt.

Sasse, Hans-Jürgen. 1993. Das Nomen – eine universelle Kategorie? *Sprachtypologie und Universalienforschung STUF* 46.3: 161-236.

Schachter, Paul. 1985. Part-of-speech systems. In Shopen, Timothy (ed.) *Language Typology and Syntactic Description* 1: *Clause structure*, 3-61. Cambridge: Cambridge University Press.

Schachter, Paul & Otanes, Fe T. 1972. *Tagalog Reference Grammar*. Berkeley: University of California Press.

Stassen, Leon. 1985. *Comparison and Universal Grammar*. Oxford: Blackwell.

Stassen, Leon. 1997. *Intransitive Predication*. Oxford: Oxford University Press.

Stassen, Leon. 2005. Predicative adjectives, WALS chapter 118.

Vogel, Petra M. & Comrie, Bernard (eds.). 2000. Approaches to the Typology of Word Classes. Berlin: Mouton de Gruyter.

Wälchli, Bernhard. In prep. Distributional parts of speech.

WALS = Haspelmath, Martin & Dryer, Matthew & Gil, David & Comrie, Bernard (eds.) 2005. The World Atlas of Language Structures. (Book with interactive CD-ROM) Oxford: Oxford University Press.

Walter, Heribert. 1981. *Studien zur Nomen-Verb Distinktion aus typologischer Sicht*. München: Fink.

Software used: R http://www.r-project.org, Python www.python.org