To classify words: Western and Indian Grammatical Approaches
Emilie Aussant

To cite this version:
Emilie Aussant. To classify words: Western and Indian Grammatical Approaches. Peter M. Scharf. Sanskrit Syntax and Discourse Structures, Jun 2013, Paris, France. 2015, Sanskrit Syntax - Selected papers presented at the seminar on Sanskrit syntax and discourse structures, 13-15 June 2013, Université Paris Diderot. <halshs-01396088>
To classify words: European and Indian grammatical approaches

Émilie Aussant

**Abstract:** This paper aims at briefly recalling the presuppositions the European and Indian grammatical approaches present regarding word-classification. From its very beginning up to the most contemporary trends in linguistics, the European approach knew essentially only one classification of words: the canonical list of eight parts of speech which comes from Dionysius Thrax and which was developed, restructured, sub-divided, etc., throughout the course of history, according to the aims of particular scholars or schools. Within Vyākaraṇa, native Sanskrit grammar, the situation is not so different: the Sanskrit grammarians generally alternated between only two classifications (the four-fold Yāskañ division of parts of speech and the two-fold division proposed by Pāṇini); some of them like Nāgeśa having even tried to reconcile the two.

**Keywords:** Europe, India, history of linguistic ideas, epistemology of language sciences

1 Introduction

Even though the analysis of language into units appears to have been essential in all traditions of language study, the classification of words consists in an activity which is neither self-evident nor homogeneous: the classifier (whether he is an individual scholar, a school of thought or a trend) has an epistemological aim (that
is to say, he has something to explain) and one must consider the regularities (hence the categories) he makes exist on this basis. Epistemological aims have varied a lot in history, from one tradition of language study to another, but also within one and the same tradition. The analysis and description of Sanskrit in the perspective of computational linguistics brings two different grammatical traditions, as well as different classifications, face to face. Without calling into question their legitimacy (my approach is rooted in the perspective of the history of ideas), I would like to recall, albeit briefly, the presuppositions these theories present. I will consider first the European approach and then the approach of Sanskrit (Vyākaraṇa perspective).

2 The European approach

2.1 Parts of speech in computational linguistics

In computational linguistics, the parts of speech which are used come, most of the time, from works realized in the field of formal linguistics. Very often indeed, taggers are statistical tools trained on manually or semi-automatically annotated data. These data were mostly annotated under the responsibility of formal linguists and according to the guidelines they elaborated on the basis of their linguistic expertise, as well as on the basis of what results from the confrontation of the system with real data. Hence, at first sight, there is no real norm: corpora can be tagged in various ways. However, some tag sets, such as the Brown Corpus

---

1I wish to thank Aimée Lahaussois, Clément Plancq and Benoît Sagot for their kind help.
tag set, the Penn Treebank tag set, and, to a lesser extent, the French Treebank developed by the University of Paris Diderot, today constitute a reference. If one looks at these tag sets, one observes that the choices made remain mainly traditional — it is even explicitly given on the website of the French Treebank, “For part of speech, we made traditional choices [...]” and, once we subtract the numerous categories and sub-categories which divide or complete each part of speech (see Table 1), and which are mainly based on inflectional and lexico-semantic information, we find the well-known list again: noun, verb, article, adjective, preposition, pronoun, adverb, conjunction, and interjection. Even tools developed for languages other than English and French, such as the Natural Language ToolKit (NLTK) tag set for four Indian languages (Bangla, Hindi, Marathi and Telugu), and the Stanford Natural Language Processing Group (SNLPG) tag set for Chinese, resort to this list. Where does this list come from? On what kind of classification criteria is it based? How can one explain its stability throughout the course of history?

2.2 Parts of speech in Graeco-Latin grammatical tradition

The list of eight parts of speech comes from the Graeco-Latin grammatical tradition, and more specifically from the Τέχνη

---

2http://www.comp.leeds.ac.uk/ccalas/tagsets/brown.html
3http://www.cst.dk/mulinco/filer/PennTreebankTS.html
4http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php
6http://nlp.stanford.edu/projects/chinese-nlp.shtml#pos
Table 1

Sub-categories of noun in the Brown Corpus

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>singular or mass noun</td>
</tr>
<tr>
<td>NNS</td>
<td>plural noun</td>
</tr>
<tr>
<td>NP</td>
<td>proper noun or part of name phrase</td>
</tr>
<tr>
<td>NR</td>
<td>adverbial noun (home, today, west)</td>
</tr>
<tr>
<td>NN$</td>
<td>possessive singular noun</td>
</tr>
<tr>
<td>NNS$</td>
<td>possessive plural noun</td>
</tr>
<tr>
<td>NP$</td>
<td>possessive proper noun</td>
</tr>
<tr>
<td>NPS$</td>
<td>possessive plural proper noun</td>
</tr>
</tbody>
</table>

Γραμματική by Dionysius Thrax (2nd-1st c. BCE). The aim of the Greek scholar was to provide, mainly for writers and poets, an overview of the Alexandrian grammatical tradition, giving a reference list for phonetics (letters-sounds, syllables) as well as the list of eight parts of speech: noun, verb, participle, article, pronoun, preposition, adverb and conjunction. This list itself is the result of developments and restructurings of a more ancient classification: that made by Aristotle (384–322 BCE) in his Poetics (chapter XX), which distinguishes four parts of speech: ὄνομα ‘noun’, ῥῆμα ‘verb’, σύνδεσμος ‘link’ (late ‘conjunction’), ἀρθρόν ‘articulation’ (late ‘article’) and which is mainly based on notional criteria (semantic, ontological, psychological, and/or logical) (Lalot 1988). The four parts of speech list was quickly enlarged and restructured on the basis of formal and semantic criteria. In the Τέχνη Γραμματική by Dionysius Thrax, each part of speech is defined, its ‘accidents’ (e.g. everything which happens to it, that is to say number, gender, inflection, diathesis and tense variations) are listed, defined and illustrated, its morphological and/or seman-
tic sub-classes are listed and illustrated (31 for the noun, 13 for the verb, 26 for the adverb, 9 for the conjunction) (Lallot 1998). What are the criteria used for this classification in eight parts of speech? Notional (see above) and formal (morphological) criteria were used very early; functional criteria (syntactical) appeared a little bit later, with the Greek grammarian Apollonius Dyscolus (2nd c. CE). But from the very beginning, parts of speech were — and are still today — classified according to (a) their meaning and/or their reference, (b) according to their form, and (c) according to their position and/or their relation to other parts of speech in sentences (Lagarde 1988). The hierarchy of these criteria, the number of their sub-categories as well as their relationship, varied greatly in history. For instance, Ramus (Pierre De La Ramée), a 16th c. French grammarian, tried to impose (in his *Scholae Grammaticae*) a model with two categories of words (lat. *voces*), using exclusively formal criteria: ‘mots de nombre’, i.e. nouns and verbs, *versus* ‘mots sans nombre’, i.e. adverbs and conjunctions.\(^7\) For authors of universal grammars (which elaborate their own categories by linking them to cognitive structures), such as Beauzée and Condillac in the case of the “Grammaire générale française”, notional criteria are much more relevant than the others (Auroux 1988: 81). As Lagarde (1988: 93) already highlights, the list of eight parts of speech, though provoking strong debate,\(^8\) crossed over the centuries and modeled grammatical consciousnesses. Every linguist must, explicitly or not, make a stand regarding this model of analysis which has become an unavoidable reference.


\(^8\)See, for instance, the debate opposing Haspelmath and Dryer, two leaders of linguistic typology. Haspelmath (2010) argues that crosslinguistic grammatical comparison cannot be based on grammatical categories, because these are language specific, while Brown and Dryer (2008) write, “While there are often ways in which languages differ from each other in terms of what word classes they have, the differences are generally small compared with the similarities.”
The judgements made with regard to this model are generally negative. Despite endlessly repeated criticisms and the periodic attempts to develop universal categories or tag sets (such as the one elaborated by Petrov, Das and McDonald in 2011), this classification remains. Auroux (1988: 82) observes quite rightly that, as a word-classification tool used in various fields of language study (language description, learning, etc.) which are not always theoretically well mastered, the list of eight parts of speech is stable, much more so than the theoretical variations suggest. Simply for communicability reasons, no theoretical change can lead to a fundamental change in the list of parts of speech. The question of parts of speech cannot be reduced to a mere matter of word classes because it is on the basis of this word-classification that European grammar was elaborated.

3 The Indian approach (Sanskrit Vyākaraṇa)

If we now move to the Indian side and, more particularly, to the field of Sanskrit Vyākaraṇa, what ways of classifying words do we find? Two models were mainly used: the nāman-ākhyāta-upasarga-nipāta classification and the Pāṇinian classification of padas.

3.1 The nāman-ākhyāta-upasarga-nipāta classification

One of the most ancient classifications of words seems to be the one in four parts: nāman, ākhyāta, upasarga and nipāta, which are traditionally translated as ‘noun’, ‘verb’, ‘preverb’ and ‘particle’. In section 13.9 of the Nirukta (N. hereafter) which is considered to be an appendix (pariśiṣṭa), while commenting on the Rg-vedic verse

---

9 See http://www.petrovi.de/data/lrec.pdf
10 Sarup (1920–1929) does not translate this section.
Speech has been measured in four steps. The Brahmins who have control over their mind know them. The three of them which are deposited secretly do not stir. The fourth [step] of the speech is that which men speak. (Bhate undated: 2, revised).

Yāska mentions different readings of the expression catvāri padāni among which the following one:

*Nāmākhyāte copasarganipātāś ceti vaiyākaraṇāḥ.*

(\textit{N.} 13.9, Sarup 1920–1929: 226)

Grammarians [consider that these four kinds of pada are]: the noun and the verb, as well as the preverbs and the particles.

Yāska himself gives this classification of padas at the very beginning of his \textit{Nirukta}:

*Tad yāni catvāri padajātāni nāmākhyāte copasarganipātāś ca tānīmāni bhavanti.*

(\textit{N.} 1.1, Sarup 1920–1929: 27)

Now, what (are) the four classes of words? They are the following: noun and verb; prepositions and particles. (Sarup’s translation)

This classification, which is found in other texts such as the Mahābhāṣya (\textit{Paspāṣā}), seems to come directly from the very early grammatical analyses elaborated in India. Such analyses aimed at deconstructing Vedic texts in their continuously recited version (\textit{samhitā-pāṭha}), where sandhis are applied, to result in a word-by-word version (\textit{pada-pāṭha}) of the same texts, where any dissociable sequence was de facto isolated. This kind of analysis implied
an accurate study of morphology and syntax, and it is most likely
during this period that the four-part classification of words was
elaborated. The fact is that it is to be found in the Rgveda-Prātiśā-
khya:

\[
\text{Nāma-ākhyātam upasargo nipātaś}\]
\[
catvāry āhuḥ padajātāni śabdāḥ;\]
\[
tan nāma yenaḥhidadhāti sattvam\]
\[
tad ākhyātam yena bhāvan sa dhātuh. (RPr. 12.5)\]

The noun, the verb, the preverb and the particle, these
are the four classes of words say the grammarians.
The noun is that by which a substance is denoted; a
verb is that by which a process [is denoted]. It is a
dhātu.

\[
\text{kriyā-vaćakam ākhyātam upasargo viśeṣa-kṛṭ;}
\]
\[
sattva-abhidhāyakam nāma nipātaḥ pāda-pūraṇaḥ.\]
\[
(RPr. 12.8)\]

The verb expresses an action; the preverb specifies
it. The noun expresses a substance; the particle com-
pletes the verse quater (pāda).

Though we have access to neither the thought nor the method of
the authors of pada-pātha texts, it is nevertheless possible to infer
from their works some of the principles they adhered to. For in-
stance, what they meant by pada, according to Bhate (undated: 4)
were isolated inflected forms (nominal, verbal), indeclinables, and
preverbs. As it appears in the first verse quoted previously, as well
as in the Nirukta, the criteria on which the distinction between the
noun and the verb is based is notional (ontological) criteria. Re-
garding the upasargas, two definitions are given by Yaśka:

- One, ascribed to Śākaṭāyana, says that upasargas do not ex-
  press meanings (na ... arthān nirāhur); they simply help the
emergence of the whole meaning by suggesting it (*karma-upasanyoga-dyotakā bhavanti*) (*N.* 1.3)

- The other, ascribed to Gārgya, presents upasargas as units expressing (*vācaka*) diverse meanings, meanings which modify the meaning of verbs and nouns with which the u-pasargas are combined (*uccāvacāḥ padārthā bhavantīti gārgyaḥ. tad — ya eṣu padārthah, prāhum ime tam — nāma-ākhyaṭayor artha-vikaraṇam*) (*N.* 1.3). The second verse quoted from the *Rgveda-Prātiṣṭhākhyā* (*upasargo viśeṣa-krīt*) seems to refer to this conception.

Concerning the particles, only one description is given in the *Nirukta*: nipātas are said to express numerous meanings (*nipātā u-uccāvaceṣv artheṣu nipatanti*): comparative (*apy upamārthe*), connective (*api karma-upasamgrahārthe*), expletive (*api pāda-pūraṇāḥ*) (*N.* 1.4).

We note that the nāman-ākhyaṭa-upasarga-nipāta classification is presented as exclusively based on notional criteria: an ontological criterion for the noun and the verb and a semantic criterion for the preverbs and the particles. This is surprising since, considering the context of the emergence of this classification, we would have expected formal criteria.

### 3.2 The Pāṇinian classification of padas

On a practical level, the *Aṣṭādhyāyī* (*A.* hereafter) provides guidelines for forming words with affixes. These affixes (*pratyaya*) are directly taught in some of the 4000 sūtras, in distinction from most of the bases with which they combine and which are either verbal roots (*dhātu*) or nominal bases. These nominal bases are:

- either generally introduced by *A.* 1.2.45 *arthavad adhātu-r apratyayāḥ prātipadikam* (“A meaningful [unit], which
is neither a verbal root nor an affix, [is called] *prātipadika*”), which is completed by A. 1.2.46 *kṛt-taddhita-samāsā-ś ca* (“Primary derivatives, secondary derivatives and compounds [are] also [called] *prātipadika*”);

• or given in lists which gather together units sharing some derivational features (cf. the *Gaṇa-pāṭha* lists (GP. hereafter);

• or directly introduced in some sūtras, such as A. 3.1.16 *bā-ṣpa-ūṣmabhīyāṃ udvamane*, which states that the (denominative) suffix *kyaṅ* is used after the nominal bases *bāspa* ‘tear’ and *ūṣman* ‘heat’ to express ‘to emit’.

Verbal and nominal bases constitute the two starting points of the derivational process which occurs over the course of Pāṇinian sūtras to generate more and more complex units up to sentences (*vākya*).11 The steps of this derivational process consist in the application of operations to these units and, generally, it is these very operations which determine the arrangement of the units in classes or categories. For instance, a unit is called *prātipadika*, that is to say, belongs to the prātipadika class because it undergoes (or must undergo) the application of several operations. Sanskrit morphology being very rich, operations to be applied are numerous, and then, the number of classes Pāṇini distinguishes is particularly high. The main classes are those shown in Figure 1 (based on that made by Bhate (undated: 7)).

At the top of the diagram, one finds the term *pada*: Pāṇini uses it to designate the general class of words, but he also gives a definition of the term (probably the most ancient to have come down to us) which clearly delimits its extension: the sūtra A. 1.4.14 states that the units which constitute the pada class are those that end

---

11The term *vākya* is used only twice in the *Aṣṭādhyāyī*, in A. 8.1.8 *vākya-āder āmantritasya . . .*, and in A. 8.2.82 *vākyaśya teḥ . . .*).
in a nominal (sup) or a verbal (tiṅ) ending, in other words, the inflected words. One notes that in the Pāñinian system, neither nā-
man nor ākhyāta are used to designate respectively the noun and the verb: neither of these terms is used by Pāñini in a metalin
guistic way. One notes too that there is an empty square on the
diagram: it stands for the nominal stems which cannot be derived,
such as vrksa, which are meaningful units but neither a verbal root,
nor a suffix, pronoun, indeclinable, primary derivative, secondary
derivative, nor compound. As far as I know, Pāñini does not give a name — other than prātipadika — to this kind of unit. Patañja-
li explains (on vt. 5 ad A. 7.1.2) that for Pāñini, nominal stems
which cannot be derived are the unādis.\textsuperscript{12} Contrary to nāman and ākhyāta, upasarga is used in the Aṣṭādhyāyī, and, here too, it designates the preverbs (cf. A. 1.4.59 upasargāh kriyā-yoge). However, it belongs to a larger class called prādi (GP. 154), which itself belongs to the nipāta class (the term is also used to designate the particles). And this nipāta class belongs to the avyaya class (‘indeclinables’, cf. A. 1.1.37 svarādinipātam avyayam) which includes also the svarādi class (GP. 254). The avyaya class constitutes a sub-category of the prātipadika class, these units being meaningful. If upasarga and nipāta are used in the Aṣṭādhyāyī with more or less the same meaning as in the Nirukta, the Pāṇinian system nevertheless establishes a difference in the level of analysis between them. However that may be, the Pāṇinian classification of padas is firstly based on formal criteria. Semantics plays a role too, but to a lesser extent; it comes as a complement (e.g. the definition of nipāta: A. 1.4.56–57 “…units which do not designate a substance (asattve)”).

3.3 Some remarks on a few other classifications

Still within the Vyākaraṇa sphere one finds mention of a few other classifications of words. I will not address the classifications which correlate linguistic categories with ontological categories, such as the jāti-guṇa-kriyā-yadrcchā classification provided by Patañjali in his Mahābhāṣya. This classification, as well as the ākṛti-kriyā-guṇa-dravya list of padārthas, actually provides a sub-classification of prātipadikas, often illustrated by the sentence gauḥ śuklaś ca ṭīthah “[the] white moving cow ḍītha”.

At the very beginning of the third kāṇḍa of his Vākyapadīya, Bāhtrhari mentions three views on the classes of padas. It

\textsuperscript{12} Prātipadikavijñānāc ca pāṇiner Siddham. Prātipadikavijñānāc ca bhagavataḥ pāṇiner ācāryasya Siddham. Unādayo ’vyutpannāni prātipadikāni. (A. 7.1.2, vt. 5; MBh. 3.241.19–21)
is very well-known that, according to Bhartrhari, padas exist only on the level of apoddhāra ‘[artificial] analysis’; in a philosophical perspective, only the sentence is relevant. But placing himself on the apoddhāra level, he says:

\[
\text{dvīdhā kaiścit padaṁ bhinnāṁ} \\
caturdhā pañcadhāpi vā; \\
apoddhrtyaiva vākyebhyaḥ \\
prakṛtipratyayādīvat. (VP. 3.1.1)
\]

Some distinguish two [kinds] of padas, [some] four [kinds] and [some] five kinds after having artificially isolated them from sentences, just as [one would isolate] the stem and the suffix, etc.

These classifications are:

- nāma (which includes nipāta) : ākhyāta (which includes upasarga and karmapravacanīya);
- nāman : ākhyāta : upasarga : nipāta;

A very similar classification is mentionned by Durga, in his commentary on Nirukta 1.1:

\[
\text{Naikaṁ padajātaṁ yathārthaḥ padam aindrāṇām iti.} \\
\text{Nāpi dve yathā subantaḥ tiṇantaś ca. Nāpi trīṇi nipātopasargāv ekataḥkṛtvā. Nāpi pañca saḍ vā yathā gatikarmapravacanīyabhedaneti. (Durga on Nirukta 1.1)}
\]

It is not the case that there is a single class of words as, ‘that which is meaningful is a word for the Aindrās’. Nor are there two [classes of words], for instance nouns and verbs. Nor are there three [classes of words] combining in one particles and preverbs [along with the two preceding ones]. Nor five or six [classes
of words], as with the distinction between gati and karmapravacanīya.

These classifications do not really improve upon the previous analyses: they can be seen as variants of the four-fold classification given in the Nirukta and the R̄gveda-Prātiśākhya.

4 Concluding remarks

From its very beginning up to the most contemporary trends in linguistics, the European approach knew essentially only one classification of words: the canonical list of eight parts of speech which comes from Dionysius Thrax. Throughout the course of history, it has been developed, restructured, sub-divided, etc. according to the aims of particular scholars or schools. The heterogeneity of its criteria and its Indo-European focus did not prevent theoreticians from analyzing speech and from proposing more and more elaborate descriptive or generative models.

Concerning the Sanskrit approach (within the Vyākaraṇa sphere), the situation is not so different. I would say — following Deshpande (1992: 14) — that “Sanskrit grammarians care for Pāṇini’s two-fold division of stems into nominal stems and verb roots only as far as the derivational procedures in his grammar are concerned.” When the discussion deals with semantics, “they generally follow the four-fold division mentioned by Yāska . . . .” The inclusion, in the Pāṇinian model, of the avyaya class in the prātipadika class seems to have been counter-intuitive for most of the vaiyākaraṇas. Some, like Nāgeśa, nevertheless attempted to reconcile the four-fold Yāskan division of parts of speech with the two-fold division proposed by Pāṇini. Commenting on the Mahābhāṣya passage catvāri śṛṅgāṇi — catvāri padajātāni nāmākhyātopapa-sarganipātāś ca, he says: nāmaśabdena subantam . . . ākhyātam
As far as the tagging of parts of speech is concerned, computational linguists working on Sanskrit are not doing anything different from their predecessors: they try to establish more appropriate models and have to hand a wonderful tool: the Pāṇinian grammar, which matches their interests particularly well. More than their 19th-20th c. colleagues who wrote Sanskrit grammars in European languages, “...they are in a position to put into practice one of the major insights of structuralist linguistics of the twentieth century, namely, that languages are best described in their own terms” (Haselmath (2010: 664) referring to Boas (1911–1922)).

Table 2
Abbreviations

| A. | Aṣṭādhyaeyī. |
| GP. | Gaṇapāṭha. |
| MBh. | Mahābhāṣya. |
| RPr. | Ṛgvedaprātiśākhya. See Régnier 1858. |

References


Bhate, Saroja. 2006. “Classifications of terms.” History of Science, Philosophy and Culture in Indian Civilization, general editor D. P. Chattopadhyaya; vol. 3.4, Philosophical Concepts relevant to Sciences in Indian Tradition, ed. by P. K. Sen, pp. 609–


d’épistémologie des sciences du langage.
Régnier, Adolphe. 1858. Études sur la grammaire védique:
Sarup, Lakshman, ed. and trans. 1920–1929. The Nighaṅtu and
the Nirukta: the oldest treatise on etymology, philology and
semantics: critical edition, translation, introduction, exegeti-
cal and critical notes, indexes and appendices. 1st ed. Lon-
don: Oxford University Press; Lahore: University of the Pan-
jab. [Reprinted with three parts in one volume: Delhi: Motilal
Sonatakke, N. S. and C. G. Kashikar, eds. 1933–1951. Rgveda-
Saṃhitā: with the commentary of Sāyaṇācārya. 5 vols. Poona:
Vaidic Samshodhan Mandal.