

The Mass–Count Distinction: Acquisition and Disambiguation

Michael Schiehlen, Kristina Spranger

IMS, University of Stuttgart
Azenbergstraße 12
70174 Stuttgart, Germany
{mike, sprangka}@ims.uni-stuttgart.de

Abstract

At least in the realm of fast parsing, the mass–count distinction has led the life of a wallflower. We argue in this paper that this should not be so. In particular, we argue, both theoretical linguistics and computational linguistics can gain by a corpus-based investigation of this distinction: Computational linguists get more accurate parses; the knowledge extracted from these parses becomes more reliable; theoretical linguists are presented with new data in a field that has been intensely discussed and yet remains in a state that is not satisfactory from a practical point of view.

1. Introduction

At least in the realm of fast parsing, the mass–count distinction has led the life of a wallflower. We argue in this paper that this should not be so. In particular, we argue, both theoretical linguistics and computational linguistics can gain by a corpus-based investigation of this distinction: Computational linguists get more accurate parses; the knowledge extracted from these parses becomes more reliable; theoretical linguists are presented with new data in a field that has been intensely discussed over the years (Löbel, 1986; Eschenbach, 1994; Oesterle, 1995; Middleton et al., 2004) and yet remains in a state that is not satisfactory from a practical point of view.

What fascinated us most in the mass–count distinction is the interaction of factors from different levels of linguistic analysis that surface in constructions which are quite frequent in the Western languages. Given a reliable classification of the pertinent constructions (Spranger, forthcoming), semantically interesting facts could be obtained relatively easily from a corpus. Such facts are not only of interest for systems that strive at language understanding, but also for those that are only concerned with getting high-quality parses; at least in German, measuring unit constructions are one of the major obstacles to parsing accuracy nowadays and present a problem that cannot be solved by generic syntactic rules. At the same time the phenomenon is subject to semantic constraints, which, if captured, could conceivably help overall system performance. In another paper (Spranger, 2005), we showed that this is indeed the case: Adding semantic information to a treebank guides a treebank parser towards finding more accurate analyses.

1.1. Structure of the Paper

The paper is organized as follows: In section 2, the traditional definition of the mass–count distinction is stated, which is later shown to be problematic from the perspective of corpus processing. Section 3 states in more detail why the definitional problems are so hard. Section 4 explains why we think that the dichotomy is still useful for automatic language processing even if it presents problems. Our claims are bolstered by a corpus study that is described in section 5. Section 6 presents an analysis of quantifying noun groups in German which sheds some light on the

mass–count distinction and opens up a way to automatically derive training data for the mass–count distinction task from a corpus. Section 7 concludes.

2. The Problem

In traditional grammar, one dimension on which common nouns are distinguished is “countability”. In the terminology that is used in traditional grammar and that we will also keep in this paper, the countable nouns are “count nouns”, and the uncountable ones are “mass nouns”. The distinction (hereafter the “mass–count distinction”) surfaces in different parts of morpho-syntax, but the most important distinguishing features are the following three:

1. Only count nouns can form a plural, or, more precisely, a plural that indicates a collection of multiple discrete entities (be they concrete or abstract).
2. Only count nouns can go together with the indefinite article.
3. Only mass nouns can be specified in the singular by real quantifiers like *much*, *little*, *enough*, *all*, *some*, or can occur in measure constructions.

3. The Mass–Count Distinction

The mass–count distinction transcends the traditional division of linguistics into morphology, syntax, semantics, and pragmatics. Thus it comes as no surprise that syntactic, semantic, and conceptual matters have frequently been mixed up in the description of the dichotomy. A most illustrative example is the term “mass–count distinction” itself: a term that is misleading in that it incautiously combines a primarily syntactic criterion (namely the usage of “count” nouns in construction with numerals) with a non-syntactic, semantic criterion (namely the ontological distinction between “mass” and discrete entities, see section 3.1.). As said before, we nevertheless stick with these terms, but will use them in a technical sense only.

3.1. Semantic Aspects of the Mass–Count-Distinction

From a semantic point of view, mass nouns and count nouns are distinguished by what they denote. Intuitively, count nouns denote a set of discrete or individuated elements,

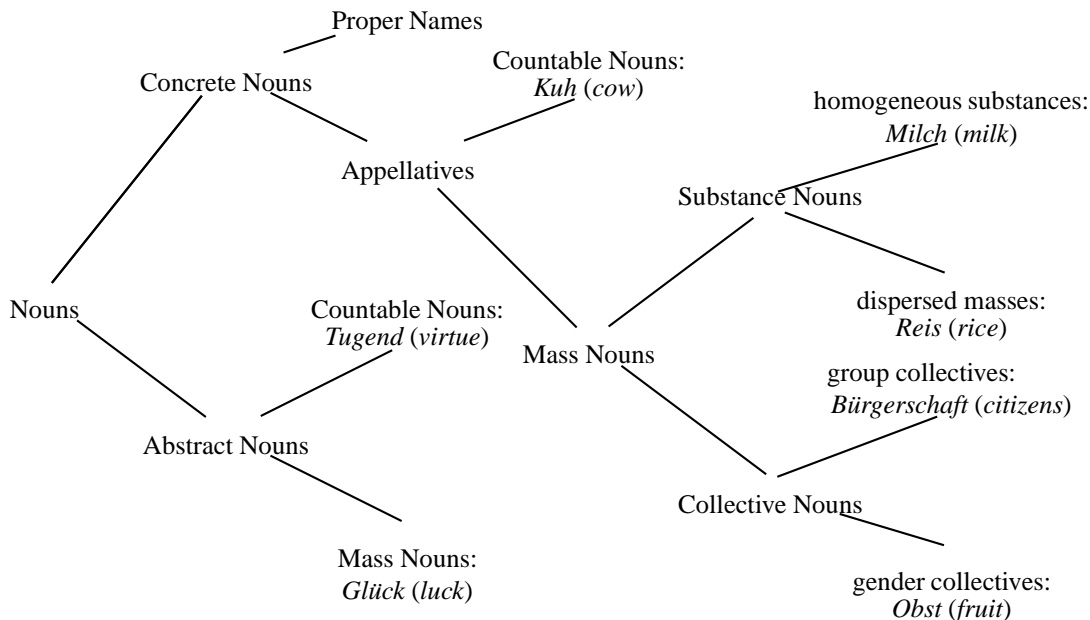


Figure 1: A Classification of Measure Constructions in German

while mass nouns denote a set without specifying how the elements are divided or individuated.

The key notion that characterizes the difference between count nouns and mass nouns has been termed “shape” (cf. (Rijkhoff, 2002)) or “bounding” (cf. (Jackendoff, 1991) or (Langacker, 1991)), i.e. the existence versus non-existence of precise limits of an entity referred to by a count noun or a mass noun, respectively. A less generic definition is not possible as bounding can apply to various domains – a cow as a physical entity has spatial limits, a beep is bound both in time and pitch, and a chapter is bounded within a written book.

To put it another way, a mass noun is semantically characterized by the fact that the parts and the sums of its denotates are also in its extension. So, for example, any sum of parts of water is again water (Quine, 1960). In contrast, the denotates of count nouns are “bounded” in some sense (Jackendoff, 1991) by e.g. a shape (in space) or a beginning and an end (in time).

3.2. The Mass–Count Distinction and Ontologies

Although a general definition in terms of their meaning is not possible, there are nevertheless “typical” regions of the semantic space (as formalized e.g. by an ontology) for mass nouns and count nouns. Mass nouns primarily refer to substances and states. Count nouns denote in particular individuals and events.

Another important distinction is that between concrete and abstract nouns. Figure 1 (Spranger, forthcoming) also shows a more fine-grained differentiation between different types of mass nouns: Certain mass nouns are like plurals in that they denote aggregates or natural kinds of discrete entities (“collectives” vs. “substance nouns”). Some of these nouns do not form plurals at all (“gender collectives” vs. “group collectives”). Among the substance nouns, some consist of clearly observable particles (“dispersed masses” vs. “homogeneous substances”). In contrast to collectives,

the sets of particles denoted by dispersed masses are neither structured nor generic.

4. The Mass–Count Distinction and its Benefits for HLT

The distinction between count nouns and mass nouns is also very valuable for HLT. It often can be utilized as a guideline for resolving syntactic ambiguity.

As far as example (1) is concerned, we can discard reading (b) as soon as we know that *Durst* is a mass noun. Otherwise, this second reading would be available - and indeed, it would be preferred by the longest-match heuristic which is often prevalent in deterministic, particularly finite-state parsing (Abney, 1997).

- (1) weil ein anderer Durst stillt
 because an other thirst quenches
 breakfasts
- (a) because another one quenches thirst
 (b) because another thirst breastfeeds

4.1. Problems of the Mass–Count Distinction

Unfortunately, the distinction is not that clear-cut in each case: in example (2), the noun *chicken* can be interpreted either as a count noun (referring to a bird) or as a mass noun (referring to meat).

- (2) I saw no chicken.

Moreover, as example (3) illustrates, it can be observed that different languages make sometimes different assumptions about the mass–count distinction.

- (3) a. Die Stadt erteilte ihm eine Erlaubnis.
 b. The town gave him permission.

In German, *Erlaubnis* is treated as count noun (otherwise it could not be preceded by the indefinite article *eine*); in

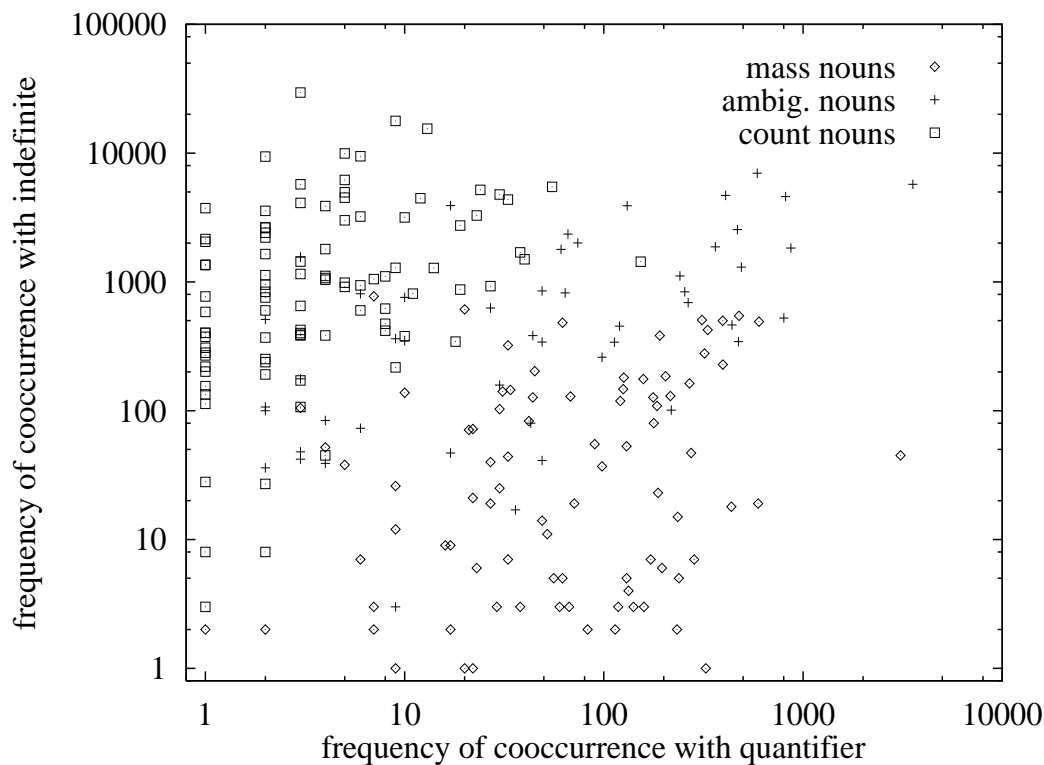


Figure 2: Space of Mass-Count Distinction

English, *permission* occurs as mass noun.

Such puzzles have led researchers to the following insight: Not the nouns should be classified in terms of count nouns and mass nouns, but their occurrences (cf. (Bunt, 1985)). In fact, mass nouns are occurrences of words which count as mass words, for the same word can occur both as mass noun and as count noun.

5. A Corpus Study

If we look at real data, it becomes clear that the traditional definition is not so much an encoding of strict regularity in language, but rather some sort of axiom that helps to express linguistic insight. In terms of the construction of HLT systems, this means that such systems can profit from it because it brings up a new dimension in the description of nominal semantics. On the other hand, it cannot really serve as a comprehensive formalization of what should be regarded as “countable”. In fact, the traditional definition leaves many nouns unclassified if rigorously applied.

We based our corpus study on a 200-million-token newspaper corpus of German from which we extracted the following three lists of nouns:

1. A list of those nouns which unambiguously occur in plural form. To this aim, we applied a morphologically enhanced noun chunker in order to make use of the agreement information provided by determiners and attributive adjectives. The chunker found 5.8 million unambiguously plural noun phrases in the corpus. We did not perform full parsing, however, i.e. we did

not make use of potential subcategorization information of the governing verbs and of subject-verb agreement, which, of course, also provides valuable sources of morphological disambiguation in German.

2. A list of those nouns which occur as heads of noun phrases that are introduced by the indefinite article. Again we made use of a chunker in extracting the list.
3. A list of those nouns that are introduced by one of the following quantifiers:
 - alle, sämtliche
 - einige, allerhand
 - lauter
 - allerlei, anderlei, beiderlei, derlei, mancherlei, mehrerlei, solcherlei, vielerlei, wievielerlei
 - zweierlei, dreierlei, viererlei, fünferlei, zwanzigerlei, vierzigerlei, hunderterlei, tausenderlei
 - dergleichen
 - viel, mehr, (all)zuviel, (eben)soviel, soundsoviel, wieviel
 - wenig, sowenig, zuwenig
 - etwas, ein (kleines) bißchen, ein wenig
 - genug, genügend, ausreichend

It should be noted that in contrast to the other quantifiers ending in *-lei*, *keinerlei* is not a mass noun quantifier.

Quantifiers that also function as adverbs (*ein bißchen*,

ein wenig, wenig, viel, derlei) were only used if they directly preceded the noun.

Another potentially relevant quantifier, the zero determiner, is unreliable as it often occurs with count nouns in fixed expressions.

As few as 6.4% of the nouns in all three lists only occurred in the third list, i.e. only 6.4% of the inspected nouns could be unambiguously identified as mass nouns. 70.1% could be unambiguously determined to be count nouns. That is to say, more than 23.5% of nouns remained ambiguous.

Furthermore, we set up a short list of 315 nouns. Since only few if any compiled lists of German count and mass nouns are available, we translated the English list given in (Leech, 1989). We manually checked it for German. Figure 2 shows the distribution for the 315 nouns in the two dimensions put up by the second and third criterion.

6. Measure Constructions

A solution that has been proposed in order to clean up the mess is the introduction of type–conversion functions. Such functions deal with the occurrences of one and the same noun in different contexts as alternatively a mass and a count noun, by positing that the noun has a primary reading which is afterwards systematically changed to the other reading.

In the realm of the mass–count–distinction, such rules have been viewed as hypothetical machines. The “Universal Grinder” (cf. (Pelletier, 1979)), for instance, can chop any object into homogeneous mass. It must have been at work in example (4a). Similarly, any mass noun can be used as count noun due to the “Universal Sorter” (cf. (Bunt, 1985)) which issues qualifications like the one in example (4b). The “Universal Packager” (cf. (Jackendoff, 1991)) pours masses into packaged and is thus responsible for cases as the one illustrated in (4c).

- (4) a. 300 g of apple
b. three wines (meaning: two sorts of wine)
c. two teas (meaning: two cups of tea)

6.1. German Count Constructions

It is interesting to note that in German many of these rules also have overt counterparts that are retrievable in a corpus. These are the count constructions, which are a subspecies of the quantifying noun groups (cf. (Spranger, 2005; Spranger, forthcoming)). Quantifying noun groups are complex nominal phrases that consist of a cardinal number, a quantity noun, and some other common noun, which we will call the quantified noun. As Figure 3 from (Spranger, forthcoming) shows, there are four subconstructions of quantifying noun groups in German (the little c’s in the figure stand for “construction”). The following listing illustrates the base constructions with the quantity nouns occurring in them:

- numeral nouns: Dutzend (*dozen*), Million (*million*)
- quantum nouns: Menge (*number*), Unmenge (*vast number*), Unsumme (*amount*), Vielzahl (*multitude*)

- abstract measuring units: Meter (*meter*), Grad (*degree*), Euro (*euro*)
- container nouns: Glas (*glass*), Tasse (*cup*), Kiste (*box*)
- action nouns (in measuring constructions): Schluck (*gulp/mouthful*), Schritt (*step*)
- relative measuring units: Prozent (*percent*)
- numeral classifiers: Stück (*piece, head*)
- shape nouns: Tropfen (*drop*), Laib (*loaf*), Scheibe (*slice*)
- unit nouns (in singulative constructions): Halm (*blade*), Korn (*grain*)
- sort nouns: Sorte (*kind*), Art (*type*)
- configuration collectives: Stapel (*pile*), Schwarm (*swarm*)
- group collectives: Herde (*herd*), Gruppe (*group*), Paar (*pair*)

Quantifying noun group constructions can be seen as the analytical counterpart to simple count constructions such as *three cows*: In languages like Chinese or Japanese, numeral classifiers are routinely inserted between cardinal numbers and pluralized nouns.

6.2. Count Constructions versus Measure Constructions

Count constructions (e.g. (5a)) are distinguished from measure constructions (e.g. (5b)) in that they do not introduce a particular dimension on which measuring is performed. Hence, they do not serve for the measurement of certain substances, but for the numerical quantification of discrete objects.

- (5) a. zehn Scheiben Brot (ten slices of bread)
b. 200 g Brot (dimension: weight)

The classification in Figure 3 indicates a systematic ambiguity in connection with container nouns: Container nouns can either occur as measuring units or as count nouns. The following example illustrates the two possibilities (6): not the plate is eaten but the soup; not the soup is smashed but the plate. Yet, syntactically there is no difference.

- (6) a. Er aß seinen Teller Suppe. (He ate his plateful of soup)
b. Er zertrümmerte einen Teller Suppe. (He smashed a plate of soup)

(Constructions like those in 6 triggered the work on measuring units in (Spranger, forthcoming), which is mainly concerned with the extraction of subcategorization frames with fast deterministic parsers. It is important for the automatic detection of selection restrictions to make a distinction between the two cases in (6). Otherwise the system would infer, as most if not all of the state-of-the-art systems do, that either plates are eaten or soups are smashed. Spranger (forthcoming) solves the problem with underspecification.)

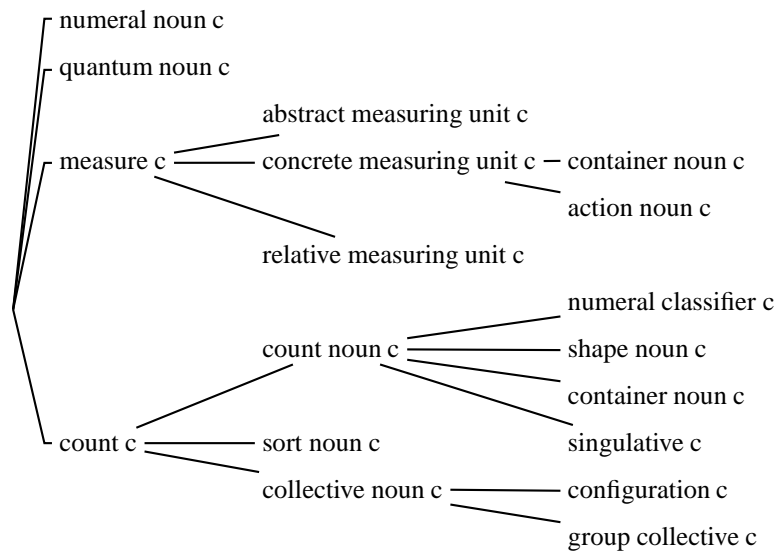


Figure 3: German Quantifying Noun Groups

6.3. Count Constructions as Type-Converters

Rather, the main purpose of count constructions is to convert mass nouns into count nouns. Again, several subclasses of count constructions can be distinguished. It is essential for the subclassification of count constructions that clearly different individuation criteria can be used for counting.

Not only the object to be counted determines the numerical value, but also the criterion used for counting: Substances can be counted if they occur in chunks (7a) or are brought into chunks (i.e. containers, (7b)). These cases correspond to the Universal Packager. Next, substances can consist of smallest particles, so-called singulatives, which can again be counted ((7c), (7d)). Finally, substances can be partitioned in different sorts, (7e), i.e. cases where the Universal Sorter was at work.

- (7) a. eine Kugel Eis (a ball of ice)
 b. ein Glas Milch (a glass of milk)
 c. ein Korn Reis (a grain of rice)
 d. ein Stück Obst (a piece of fruit)
 e. eine Sorte Reis (a kind of rice)

7. Conclusion

Using the criteria we have proposed, a list of nouns can be extracted from a corpus that classifies them as mainly mass or mainly countable. Such a list is needed in several HLT applications.

Furthermore, we have sketched an approach to cope with the type-conversions between mass and countable nouns, that are so frequent in our data. The extracted classification can be used in order to retrieve occurrences of mass nouns in which they are used as countable nouns. In such cases, a type-conversion rule like the Universal Sorter or

the Universal Packager must have been at work. The selection between the Universal Sorter and the Universal Packager is facilitated by further knowledge about typical contexts: Once we have decided for the Packager, the usage in count constructions tells us the most probable container or shape noun, *tea*, for instance, is typically served in a *cup*, and *vanilla ice-cream* is distributed in *balls*).

For distinguishing between the Universal Sorter and the Universal Packager, the local syntactic context may be important.

8. References

- Steven Abney. 1997. Partial Parsing via Finite-State Cascades. *Journal of Natural Language Engineering*, 2(4):337–344.
- Harry C. Bunt. 1985. *Mass Terms and Model-Theoretic Semantics*. Number 42 in Cambridge studies in linguistics. Cambridge University Press.
- Carola Eschenbach. 1994. Maangaben im Kontext - Variationen der quantitativen Spezifikation. In Sascha W. Felix, Christopher Habel, and Gert Rickheit, editors, *Kognitive Linguistik*, pages 207–228. Westdeutscher Verlag, Opladen.
- Ray Jackendoff. 1991. Parts and Boundaries. *Cognition*, 41:9–45.
- Ronald W. Langacker. 1991. *Concept, Image, and Symbol. The Cognitive Basis of Grammar*. Mouton de Gruyter, Berlin.
- Geoffrey Leech. 1989. *An A-Z of English Grammar and Usage*. Edward Arnold.
- Elisabeth Löbel. 1986. *Apposition und Komposition in der Quantifizierung. Syntaktische, semantische und morphologische Aspekte quantifizierender Nomina im Deutschen*. Number 166 in Linguistische Arbeiten. Max Niemeyer Verlag, Tübingen.
- Erica L. Middleton, Edward J. Wisniewski, Kelly A. Trindel, and Mutsumi Imai. 2004. Separating the chaff from the oats: Evidence for a conceptual distinction be-

- tween count noun and mass noun aggregates. *Journal of Memory and Language*, 50:371–394.
- Jürgen Oesterle. 1995. *Syntaktische und semantische Aspekte von Makonstruktionen im Deutschen*. Ph.D. thesis, Centrum für Informations- und Sprachverarbeitung, LMU München.
- Francis Jeffrey Pelletier. 1979. *Mass Terms: Some Philosophical Problems*. Reidel, Dordrecht.
- Willard Van Orman Quine. 1960. *Word and Object*. The MIT Press, Cambridge, Mass.
- Jan Rijkhoff. 2002. *The Noun Phrase*. Oxford University Press.
- Kristina Spranger. 2005. Some Remarks on the Annotation of Quantifying Noun Groups in Treebanks. In *Proceedings of the 6th International Workshop on Linguistically Interpreted Corpora (LINC-2005)*, pages 81–90.
- Kristina Spranger. forthcoming. *Combining Deterministic Processing with Ambiguity Awareness – The Case of German Quantifying Noun Groups*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung.