

# Some Words on *Word*

Maarten Janssen & Albert Visser

April 11, 2002

## 1 Introduction

In many disciplines, the notion of a *word* is of central importance. For instance, morphology studies *le mot comme tel, pris isolément* (Mel'čuk, 1993 [74]). In the philosophy of language the word was often considered to be the primary bearer of meaning. Lexicography has as its fundamental role to catalogue the words of a language. Thus, there has been a lot of work on words and on the nature of words. Examples are 'Word and Object' (Quine, 1960), 'Het Woord' (Reichling, 1989), and 'on the Definition of Word' (di Sciullo, 1988). All these works, however, focus primarily on the semantic content of words. The metaphysical questions about the nature of words have received relatively little attention. Probably the most thorough analysis of the notion is to be found in (Lyons, 1995), who makes a categorisation of different notions of 'word'. A refreshing new look at the metaphysics of words can be found in (Kaplan, 1990).

In this article, we will attempt an analysis of the notion of a 'word' which does justice both to the demands of philosophy and of linguistics. We do not aim to be exhaustive here. Such an endeavour would merit a book or a series of books. We rather aim at providing something like a map of the various issues.

Given the delicate nature of the subject, it should hardly be surprising that one needs a rich ontology to account for all the subtle problems. This article will try to provide as detailed an ontology as possible given its limited length, and to provide a proper nomenclature for the many different aspects of words. Where possible, we will follow established terminology, starting from the definitions given by Lyons (1995).

Throughout this article, we will try to use a coherent notation for the different entities that will be introduced: strings or orthographic words will be put in courier, /utterances/ or /phonological words/ in IPA between slashes, wordforms in sans-serif, **lexemes** in bold-face and words-qua-basic-primitive-things will be underlined. Only *emphasised* and 'quoted' styles will be used in their normal, sloppy way.

## 2 Words and Media

Aristotle in the beautiful opening passage of *ΠΕΡΙ ΕΡΜΗΝΕΙΑΣ* (or: De Interpretatione, see e.g. (Aristotle, 1966)) makes a threefold distinction concerning words. He says that what we utter with the voice stands for or symbolises what the soul experiences. What is written stands in its turn for the spoken word. Clearly, Aristotle draws attention to an important aspect of words.

Words are realised in different media: they come as spoken words and as written words. Certainly, words do evoke emotions and mental pictures and these are an important aspect of how words function. However, we disagree with Aristotle both on the asymmetry between the spoken and the written and on the centrality in semantics of the experiences of the soul.<sup>1</sup> Since the second disagreement touches on the question of the nature of meaning, one of the main themes of the philosophy of language, there are extensive discussions of the matter by such different philosophers as Gottlob Frege (see (Frege, 1988) and (Frege, 1976)) and Ludwig Wittgenstein (see (Wittgenstein, 1971)). We refer the reader to these discussions and do not go into the issue here.

Let's discuss the first disagreement. It is understandable why Aristotle thought like he did at a time when reading-out-loud was a much more common practice than it is now. Probably, many people couldn't understand a written text without simultaneously pronouncing it. (See (Manguel, 1996), chapter *The Silent Readers*.) Written words appeared as instructions to produce certain sounds.<sup>2</sup> Hearing these sounds produced understanding in the soul. Secondly, the spoken word did indeed precede the written word in history. One may well speculate that the emergence of language is intimately connected with the possibilities of the human vocal organs. Thirdly, the spoken word is learned first in the development of the individual.

However, by now the written word quite generally functions without the help of pronunciation. Of course, in an appropriate context, an instruction on how to pronounce it, will be associated to a written word, but conversely in a symmetric way an instruction on how to write it, may be associated to a spoken word. Moreover, some words undoubtedly emerged in the written medium first. E.g. the Dutch author Simon Carmiggelt introduced the verb epibreren ('to epibrate'). It is a word denoting senseless administrative actions.<sup>3</sup> Now we can very well imagine that Carmiggelt never spoke the verb for himself when he wrote the piece in which he introduced the word; we can imagine that the word was only pronounced loudly by his readers much later.<sup>4</sup> So here we have a possible case of a written word preceding its spoken version. On the other hand, surely, Carmiggelt must have had a pronunciation in mind, so epibreren was never separated from at least the *possibility* of pronunciation. So this argument, is perhaps not air-tight. We may however, imagine a language that was always written, never spoken and never intended to be spoken. We submit that this language would contain words. Hence, being pronounced is not necessarily the primary form of utterance of words. Sign languages such as ASL (American Sign Language) provide an example of a language that is primarily not pronounced.

Finally, some words seem to live primarily in the written medium. Consider the word like Z. Is it a word that can also be spoken? Well, when we

<sup>1</sup>The primacy of the spoken was also asserted by de Saussure. See e.g. (de Saussure, 1972).

<sup>2</sup>One may wonder if the Aristotelean doctrine does not imply that written words are also words but words that are different from the associated spoken words, specifically words with a different meaning. The written word means an instruction to produce a spoken word and the corresponding spoken word means whatever it means.

<sup>3</sup>Later a homophonic and homographic word denoting a non-existent isle on which an (existing!) group of poets is supposed to live, was introduced.

<sup>4</sup>In fact, the case is even subtler: the word was introduced, in written text, *as spoken by a fictional character*.

read an English text containing the string 7, we will utter /'sevən/. But with this utterance there also corresponds, in English, the string seven. So if the spoken utterance really is an utterance of 7, then it follows that 7=seven. But in a similar way we could argue that 7=sieben, where sieben is the German word for seven. Ergo seven=sieben. Quod non. There are two ways out of this 'paradox'. First we can claim that the string 7 as occurring in English text realises the word  $Z_{\text{engl}}$  and that 7 as occurring in German text realises  $Z_{\text{germ}}$  and that  $Z_{\text{engl}} \neq Z_{\text{germ}}$ . The other road is to say that 7 as used in English written text and in German written text always realised the single word 7. However, when we 'pronounce' it we substitute the pronunciation of a different, closely related, word, /'sevən/ in English and /'zi:bən/ in German. If we go for the second way, then we have here an authentic example of a word peculiar to writing. Surely the second way holds some attraction. Are not Arabic numerals a wonderful invention now used by nearly all civilised people?

In a curious contraposition to Aristotle, we quote footnote nr. 5 of (Tarski, 1944). (This paper has been reprinted many times e.g. as (Tarski, 1994).)

For our present purposes it is somewhat more convenient to understand by 'expressions'. 'sentences', etc., not individual inscriptions, but classes of inscriptions of similar form (thus, not individual physical things, but classes of such things).

So Tarski does not seem to count spoken expressions as expressions at all.<sup>5</sup> He also seems to exclude expressions projected on a screen and expressions appearing on a computer screen. The reason for his stress on written language, is somewhat understandable since Tarski was primarily thinking of printed mathematical texts. His insistence on inscriptions is understandable since things like overhead projectors, computers and neon letters were less ubiquitous in 1944. (But, surely, *the spoken word* existed ...)<sup>6</sup>

We want to oppose to Aristotle's (and Tarski's) ideas, a picture of what words are. Words are, in a sense, abstract objects that have realisations in different media. Here 'abstract' is not meant in the sense that the word is supposed to be an 'abstraction' from concrete instances or something like that. Also we do not intend that words are atemporal objects like numbers or forms. To the contrary we think that words are essentially historical objects. We just want to urge that words should not be identified with anything simple like a sound type or whatever. The same word may occur in different media.<sup>7</sup> In fact, we think that the usual restriction to two media is just based on a historical accident. Words may acquire new realisations in other media. E.g. words may be stored in a computer. This storage has a certain correspondence with the written version, however it *is* not a written version of the given word. Further the word is also not something like the 'sum' of the types of its realisations in different media, if these types are taken to be constituted by some equivalence relations like equiformity. *Was das Wort am inneren zusammenhält* is not anything neutral. E.g. whether a sound carries a word depends not only on the

<sup>5</sup>He also considers tokens expressions as 'physical things' and type expressions as classes of such things 'of similar form'. We hold him wrong on both counts: token expressions are not in any strict sense physical things and, whatever token expressions are, what classifies them together into type expressions is *not* similarity of form. We will discuss these matters further in section refytyo.

<sup>6</sup>For further discussion of spoken versus written words, see e.g. (Katamba, 1994), section 7.6.

<sup>7</sup>Note that the case of numerals speaks against this idea as true for *all* words and all media. Evidently a more careful discussion of these issues is needed. Will will not attempt this here.

internal qualities of this sound—even if these are not irrelevant—but also on the context in which the sound is produced. We think it is plausible that this notion of context must have intentionality built into it. In other words, we do not believe that reductionist theories of wordhood are likely to be true.<sup>8</sup>

We end this section by a brief discussion of the written and spoken realisations of words. The written realisation of a word is a sequence of letters, also called a *string* or an *orthographic word*. We will use *string* and *orthographic word* in a technical sense, *not* for ‘the word as it occurs in its written version’ or ‘the written word’, but for the written aspect of the word considered *an und für sich*. E.g. the string *verflucht* may in one context be a realisation of the Dutch word for *smell of paint* and in another context a realisation of the German word for *damned*.<sup>9</sup> In yet another context, say when typed by a monkey playing with a typewriter, *verflucht* might support no word at all: it would be just a meaningless string.

The identity of strings is at first sight straightforward: strings that have equal letters on equal places are identical. We think that the difficulty of the question of letter and string individuation has been somewhat underestimated. We will briefly touch on the various problems in section 4.2.

We distinguish strings as *types* and strings as *tokens*: every use of the on this page (for instance) is in some way a different string (token string), although they are also instances of the same string (type string). There is a third notion *occurrence* that plays a role both at the level of token and of type. Both the type-token distinction and the notion of occurrence play a role not just for orthographic words, but also for many other ‘kinds’ of words (better: *aspects* of words) and, of course, also for other syntactic entities like sentences. We will return to the matter of types, tokens and occurrences in section 4.2.

We will call the spoken unit corresponding to a word a *phonological word*. Since a word is hardly ever pronounced exactly the same way twice, and can have substantial variation over dialects, the identity of phonological words is a delicate matter. It is commonly identified with a certain ‘prototypical’ or commonly accepted pronunciation, often represented in the International Phonetic Alphabet (IPA) in dictionaries. (These representations are also words, but different ones from the words whose pronunciations they represent.) Also phonological words come in types and tokens, token versions are often called *utterances*<sup>10</sup>.

We give a brief overview of separating examples concerning *word*, *orthographic word* and *phonological word*. We already saw the example of the string *verflucht* that can be used to realise different words.<sup>11</sup> So the same string may support different words. Conversely different strings and sounds may be associated to one word. The English word medieval has two different ways of spelling: either as *medieval*, or as *mediaeval*. In the same fashion, there are also two ways of pronouncing it: either as /mediːvɪl/ or as /mi:diːvɪl/. Given the definitions of orthographic and phonological words, this means that there are both two distinct orthographic words, and two distinct phonologi-

<sup>8</sup>For a discussion of the metaphysical status of words, see (Kaplan, 1990).

<sup>9</sup>An example from (Corstius, n.d.).

<sup>10</sup>‘Utterance’ has a different connotation in the philosophical literature, as will be discussed later on.

<sup>11</sup>Note that the different words associated to *verflucht* carry different syllabifications: *verf · lucht* for *smell of paint* and *ver · flucht* for *damned*.

cal words associated with medieval. However, these differences do not cause there to be two words medieval. Something else does. There are actually two words medieval<sub>adj</sub> and medieval<sub>noun</sub>: the first is an adjective and the second a noun.

### 3 Word-Forms and Lexemes

In this section, we discuss some abstractions of the notion word that are important in linguistics. These abstractions will have *stipulated meanings*.

Our first abstraction is *wordform*. The notion of *wordform* provides an abstract model that cuts across spelling and pronunciation. A word-form can have spelling and pronunciation variations. But conversely, different word-forms can also have the *same* spelling, for instance the Dutch word-forms band<sub>1</sub> (band) and band<sub>2</sub> (tyre) are pronounced respectively as /bænd/ (band) and /bant/ (tyre) but are both written as band. The distinguishing feature between these two word-forms is thus the pronunciation. Such distinct word-forms that have the same spelling are called *homographs*, whereas word-forms that are pronounced in the same way (like lesson and lessen) are called *homophones*.

But two word-forms can even be homophones *and* homographs at the same time, while still being distinct. The word-class distinguishes the noun hamer from the (related) verb spelled and pronounced the same way. In a gender-sensitive language like Dutch<sup>12</sup>, you find differences between a neutral word bal (ball; party) and a male word bal (ball; sphere). Though the notion of a word-form is often treated as a well-defined concept, there does not seem to be a clear definition of its identity criteria. Here is an attempt at a definition.

**Word-Form** A *pre-word-form* is given by a spelling-cum-syllabification and a pronunciation.

A *word-form* is given by a number of pre-word-forms plus a wordclass plus a gender (if applicable).

We can *model* the notion of word-form as a tuple of a set (viewed disjunctively) of pre-word-forms, a wordclass and, possibly, a gender. With the word medieval<sub>noun</sub> we may associate the word-forms:

$$\langle \{ \langle \text{me} \cdot \text{di} \cdot \text{e} \cdot \text{val}, / \text{medi}:'\text{vl} / \rangle, \langle \text{me} \cdot \text{di} \cdot \text{e} \cdot \text{val}, / \text{mi}:\text{di}:'\text{vl} / \rangle, \langle \text{me} \cdot \text{di} \cdot \text{ae} \cdot \text{val}, / \text{medi}:'\text{vl} / \rangle, \langle \text{me} \cdot \text{di} \cdot \text{ae} \cdot \text{val}, / \text{mi}:\text{di}:'\text{vl} / \rangle \}, \text{count noun}, \text{male} \rangle$$

Since, set-theoretical notation does not make for pleasant reading, we replace

<sup>12</sup>Dutch does not heavily use gender.

it by a more convenient box format. The above representation thus becomes:

me · di · e · val	/medi'i:vl/
me · di · e · val	/mi:di'i:vl/
me · di · ae · val	/medi'i:vl/
me · di · ae · val	/mi:di'i:vl/
count noun	
male	

The scheme for medieval<sub>adj</sub> becomes:

me · di · e · val	/medi'i:vl/
me · di · e · val	/mi:di'i:vl/
me · di · ae · val	/medi'i:vl/
me · di · ae · val	/mi:di'i:vl/
adjective	

We allow for the possibility of spelling and pronunciation to be linked here. If we ignore this possibility, we can represent a word form by a tuple of a set of spellings-cum-syllabification, a set of pronunciations, the word class and, if necessary, the genus. E.g. our word-form for medieval<sub>noun</sub> would become:

me · di · e · val
me · di · ae · val
/medi'i:vl/
/mi:di'i:vl/
count noun
male

Arguably our present representation is not quite correct. E.g. the Dutch word idee can be both male and neutral. This is easily repaired by allowing more than one gender.

Note that the word-form is not identical to the word: it is a standardised way to represent certain crucial features of a word.

We can build an even more abstract notion. We say that *is* and *are* are different word-forms, but that still, in some way, they are the same: they are *morphologic expansions* or *inflections* of the same *word-expression*. Thus, the rows in table 1 contain different word-forms of the same word-expression.

book	books	book's	books'	
mouse	mice	mouse's	mice's	
great	greater	greatest	greatly	
bad	worse	worst	badly	
look	looks	looked	looked	looking
go	goes	went	gone	going
be	am	is	are	was   been   being

Table 1: Inflection in English

For English, every word-expression has only a few word-forms. However, if you consider a much more inflectional language like Czech, word-expressions can have lots of inflection, an example of which is given in table 2.

	<i>Sing</i>	<i>Plur</i>	<i>Pl.Masc.Sing.</i>	<i>Gen.Fem.Sing.</i>	<i>Fem.Plur</i>
<i>Subj.</i>	bratr	bratři	bratřív	bratrova	bratrovy
<i>Obj.</i>	bratra	bratry	bratrova	bratrovu	bratrovy
<i>Gen.</i>	bratra	bratřú	bratrova	bratrovy	bratrových
<i>Dat.</i>	bratru	bratrum	bratrovu	bratrově	bratrovým
<i>DL</i>	bratrovi	bratrech			bratrových
<i>Instr.</i>	bratrem	bratry	bratrovým	bratrovou	bratrovými
<i>Voc.</i>	bratře				

Table 2: Inflection of **bratr** (brother) in Czech

By convention, every word-expression has a *citation-form*, depending on the language and the word-class. The citation-form for verbs in English is the infinitive, for adjectives in Czech it is the singular male form. These citation forms are the *headwords* of definitions in dictionaries. However, headwords of dictionary definition are more generally citation forms of *lexemes* rather than word-expressions. Lexemes are strictly speaking not words but *phrases*, since there are also *multi-word lexemes*: semantically inseparable units consisting of more than one word, such as **pass muster** and **food poisoning**. Also idiomatic constructions such as **red herring** (in its non-literal meaning) are considered multi-word lexemes.

The lexemes can be said to be represented by their citation- or entry-forms: “The lemma, when used as an entry-form, conventionally represents all the inflected forms of the unit, **umbrella** for umbrella an umbrellas, **take** for take, takes, taking, taken, and took, or **go** for go, goes, going, gone, and went: inflected forms are normally all treated together in the same entry, under the same entry-form.” (Béjoint, 1994 [192]).

Notice that the same word-form can represent different lexemes; For instance, the French word-form *juste* (adj.) can either be expanded as *juste-justement-justesse*, or as *juste-justement-injust-justice-injustice* (Messelaar, 1990 [39]). Since there are no formal differences between these two word-forms *juste*, they are the same word-form representing completely different lexemes.

## 4 Strings

Strings are not words. However, words are sometimes realised as strings. Moreover, the notions *type*, *token* and *occurrence* are applicable both on the string and on the word level. In this section we will study the notion of a string. We will explore sameness of strings and the fundamental notions of *type*, *token* and *occurrence*.

### 4.1 Sameness of Strings

Two strings are the same if they consist of the same characters in the same order. Of course, we want to allow a certain variation in a string, e.g. change of font, which preserves a string as the same. ‘Mooi weertje vandaag’ is the same string as ‘Mooi weertje vandaag’. What causes this sameness? The forms of the letters play an important role in identifying the letters and it is clear that there

are similarities in form across fonts. However, form cannot be the final answer. Sameness of form is neither necessary nor sufficient for sameness of string.

### **Equiformity Is Not Sufficient.**

The Greek capital string 'EPMHNEIA' would be spelled as 'ερμηνεία' in lowercase. The roman string 'EPMHNEIA' would be spelled in lowercase as 'epmh-neia'. So Greek 'EPMHNEIA' is different from roman 'EPMHNEIA', even if they share the same form. These different strings can even be produced by the same actions on the keyboard.

Leslie Lamport writes:

*If the uppercase Greek letter is the same as its Roman equivalent, as in uppercase alpha, then there is no command to generate it.* See (Lamport, 1994), p41.

This is deliciously confused. Presumably 'equivalent' does not mean sameness of form, since otherwise the antecedent of his implication becomes tautological. So it will mean either historical relatedness or being associated to a similar sound. However Greek 'H' and Roman 'H' are not equivalent in this sense. Still there is just one command to generate them. Secondly, of course, *there is* a command to generate capital alpha. It's just the same command as the one for capital a.

What makes the Greek string different from the Roman one is just difference in intention, difference in context. In our example this difference is provided by an explicit statement of ours to the effect that the one is Greek and the other is Roman. (Embarrassing question: could we have lied about this?) Usually, we know these things by contextual inference. E.g. when we wrote earlier in the paper "Aristotle in the beautiful opening passage of ΠΕΡΙ ΕΡΜΗΝΕΙΑΣ ...", you understood that the substring "EPMHNEIA" was Greek.

A second example of the miraculous workings of context is provided by boustrophedon. Boustrophedon ('as the ox plows') is a way of writing where lines are alternately written from left to right and from right to left. Going from right to left the letters are mirrored along the vertical axis. Thus, the form 'd' in from left to right direction stands for the same letter as the form 'b' in right to left direction. The same form 'd' stands for the fourth letter of the alphabet in left to right direction and for the second in right to left direction.

Thirdly, a form may occur accidentally, e.g. an island shaped like an A. This island is not an A.

Summarizing, one could say: what makes a letter a letter is never just form, it is always also context. Note that this objection works as well against a family resemblance theory of what makes a shape to an instance of a letter.

### **Equiformity Is Not Necessary**

If each token of a letter would have a completely different form, if there were no stability of form through time at all, then letters would not be very useful. Even so much is not quite true. E.g. we could imagine that these changes would follow fixed rules—a bit like boustrophedon. Anyway, we are proceeding from the assumption that letter-tokens are visual things, something like recognisable contours. Thus, *form* plays a necessary role. However, there is

no natural notion *same form* to capture sameness across fonts. E.g. one should simply know that  $\text{ƀ}$  and  $\text{z}$  are the  $k$  and  $z$  of the fraktur font.

Could we say that a token of a letter is identical with the contour of a visible thing, even if the letter-qua-type is not a form? We postpone the discussion of this issue to our discussion of types, tokens and occurrences below.

### Non-preservation

It may help us focus on the problem, if we consider some cases that fail to produce the appropriate sameness of letter or string even if some sameness is there.

A primary example is provided by the familiar decoding puzzles in magazines. These puzzles involve a simple 1-1 substitution of icons for letters. Similarly, we may change an alphabetic string in MS Word by a 'font'-switch into a string of Zapf Dingbats. Such changes are 1-1 and automatic. In the puzzles the icons are intentionally connected to the corresponding letters. So why do we not say that these icons *are* the corresponding letters? If we can identify 'z' and '3', then why not also 'n' and '■'? The answer is probably that (i) there is no shared intention to use these icons as letters *in a sustained way*, and that (ii) the possible intention to use them as letters would not be reasonable because of the utter lack of similarity with the forms usual for these letters in the local environment.

An easier example is provided by Japanese strings. Where English has only 26 characters, Japanese has some 5000 Kanji characters. But also, Japanese has three independent writing systems: kanji, hiragana and katakana. Kanji is the Japanese version of Chinese characters, whereas hiragana and katakana are both syllable-based writing systems. Often, the same word can be written either in kanji or in hiragana. An example is given below, where the Japanese word for **beach** (/hama/) is represented both as a single Kanji symbol, and as the syllabic hiragana symbols for /ha/ and /ma/ in sequence.

浜            はま            /hama/            beach

Superficially this is a bit like using another font. But kanji and hiragana aren't simply different fonts: they are different writing systems, based upon a completely different principle. So the difference between the two is easily shown: two different kanji signs, with the same Japanese pronunciation, will be transcribed into the same string in hiragana. Since from the kanji perspective these two strings should be different, and from the hiragana perspective they have to be identical, the conclusion is that the kanji and its hiragana transcription are distinct strings.

## 4.2 Types, Tokens and Occurrences

The distinction between types and tokens was introduced by Charles Peirce in 1860, as part of his pragmatic theory. It was originally introduced as a tripartition, with *tones* as the third class:

A common mode of estimating the amount of matter in a MS. or printed book is to count the number of words. There will ordinarily be about twenty *the's* on a page, and of course they count as twenty words. In another sense, of the word "word" however, there is but one word "the" in the English language; ... Such a definitely significant Form, I propose to term a *Type*. ... a Single object or thing which is in some single place at any one instant of time ... such as this or that word on a single line of a single page of a single copy of a book, I will venture to call a *Token*. An indefinite significant character such as a tone of voice can neither be called a Type nor a Token. I propose to call such a Sign a *Tone*. In order that a Type may be used, it has to be embodied in a Token which shall be a sign of the Type, and thereby of the object the Type signifies. I propose to call such a Token of a Type an *Instance* of the Type. (Peirce, 1906 [423])

We feel that the notion of "occurrence" should be mentioned here too. The time is ripe for a more elaborate philosophy of types, tokens and occurrences. In this paper, we will not try to develop such a philosophy in full. We will just dwell on certain salient points.

The type, token, occurrence distinction is applicable to letters, strings, words, sentences, etc. . We will mainly discuss it for letters and strings here.

Is there a token-token identity between a token letter and something like a token contour? The answer depends on quite broad matters of one's general philosophy concerning individuation. If one thinks, as we tend to do, that, at least at a given moment, 'the kind of a thing is built into the thing' in the sense that we cannot coherently imagine the same entity at the same moment in time having a different kind, then, of course, the answer is an easy *no*. Since the letter qua kind is not equal to anything neutral—like a form—a fortiori its instances cannot be equal to realised forms. If one thinks that the same thing may instantiate several non hierarchically ordered kinds, then the answer could be *yes*.

We will not try to resolve these metaphysical matters within the modest space of this paper. Rather, we will have a look at examples and see what they tell us about the way we think of token letters.

Tokens of letters are often thought of either as physical objects or as closely associated to physical objects, e.g. they could be boundaries or contours of physical objects. Surely, some token letters, such as the letters of my copy of the L<sup>A</sup>T<sub>E</sub>X user's guide lying here beside me, are precisely that. In fact the association of letters with stable physical things was precisely what constituted their initial usefulness: *verba volant, scripta manent*.

Consider e.g. letters on the computer screen. If one places the cursor in front of a letter and types a space, the letter will move to the right. Similarly one may pick up the displayed document and put it on a different place on the screen. If one looks at a menu on one of our MacIntoshes and points with the mouse at one of the commands in the menu, the black letters against a white background become white letters against a coloured background. In all of these cases are we looking at the same tokens? Our talk of movement and change seems to suggest so. Of course, there are *not really* physical objects moving or changing. Certain actions result in certain instructions to the computer which change the internal state. The computer's changed state causes a change of the screen. One could e.g. in principle make the letters on the screen move faster than

light, faster than any physical object can. Similar examples, can be given using or letters projected on a screen by an overhead projector or ‘moving letters’ on an electronic billboard.

Of course, we could say that tokens of letters are only ‘instantaneous objects’, thus cancelling our talk of movement and change. However, we feel that this option is the less natural one. (How short should the instant be? If it is too short, you never see a token but only a sequence of tokens.) So we are inclined to say that token letters are visible contours on a suitable surface whose transtemporal identity is constituted by continuity and probably context. Token letters can be, but need not be directly associated to physical objects having the relevant contour.

Here is a second class of examples. The usual way to introduce the difference between types and tokens is to ask how many letters there are in e.g. *distinction*, with as its possible answers either 8 or 11, depending on whether you count types or tokens. But now look at the first and the second word in the figure on the right below: how many tokens of the letter ‘S’ does that contain? If the answer is *one*, than of which type are the tokens ‘P’ and ‘b’ on the other two examples instances? And this cannot be ascribed to indefiniteness: the ‘P’ in the leftmost example is definitely a *Rho*, and equally definitely the roman ‘P’.

C U E P O Σ I D O	a   λιb e r t	FIRST E C O N D
----------------------------------	------------------------------	--------------------------------

Let’s first concentrate on the ‘P’ in leftmost figure. There are two options.

1. There is one token that *is* the different letters *Rho* and roman *P* at the same time. This view would well fit with the theory that allows token identity of a letter with a contour. So not every ‘P’-shaped contour constitutes a *Rho*, but if it does, this contour *is* this *Rho*. We still have the notion of occurrence to get at the desired difference: we may hold that there are one occurrence of *Rho* and one of roman *P* here and that these occurrences are different.
2. There are two tokens, one of *Rho* and one of roman *P* here. They just happen to occupy the same space. This view coheres well with the philosophical conviction that the sort is built into the object.

As we already indicated we would favour the second option. However it is important to realise that the notion of occurrence—which is intrinsically linked to the notion of text—is probably more important than the notion of token. So our precise choice here will probably not have very heavy consequences for the philosophy of language.<sup>13</sup>

<sup>13</sup>Perhaps the choice between (1) and (2) is only a matter of taste. Here is an ‘argument’ to strengthen this impression. Given (1)-tokens we can define (2)-tokens as those contours that realise

What about the rightmost figure? Here it seems reasonable to say that we do have one letter ‘s’. However, we *do* have two occurrences.

We can repeat our example above for words. the following story is philosophical folklore —as far as we know.

A woman is talking to her Icelandic friend, while her Danish husband walks into the room and gives her a cup of tea. The woman utters /tak/, thus simultaneously thanking her husband (it means *thanks* in Danish), and finishing the conversation with her friend (it means *bye* in Icelandic).

Let’s assume that our man obtained the effect described intentionally. In this case we would prefer to say that he uttered two words employing one sound —what splendid efficiency. The other option would be to say that he uttered one word that was simultaneously an Icelandic and a Danish word. If one follows this last option one could still say that the man engaged in speaking to different text and that the one word carries different *occurrences* of words.

Finally, we turn to the notion of occurrence. An important paper on the notion of occurrence is Linda Wetzel’s (Wetzel, 1993). As Wetzel correctly points out occurrences cannot be identified with tokens: there are three occurrences of *Macavity* in the type *Macavity, Macavity, there’s no one like Macavity*. If occurrences were tokens, tokens would occur in a type. Wetzel does not distinguish words and strings, nor does she distinguish sentences and strings. We first discuss strings to return to the more general question later. So the question is: what is an occurrence of *Macavity* in \*: *Macavity, Macavity, there’s no one like Macavity*? Wetzel correctly points out that occurrence is a positional notion: a string occurs in a string at a certain place. She proposes to analyse occurrences as triples  $\langle n, \sigma, \tau \rangle$ , indicating the  $n$ -th occurrence of the string  $\sigma$  in the string  $\tau$ . Thus the second occurrence of *Macavity* in \* would be  $\langle 2, \text{Macavity}, * \rangle$ , etc. Note that:

1. Wetzel provides a *modelling* for the notion of occurrence. We need not really believe that an occurrence is an ordered triple. Lots of alternative modellings suggest themselves.
2. Already at this level there are ambiguities. There is a difference between a sentence-as-string, where the blank is treated as just another letter, and a sentence-as-sequence-of-strings. Consider e.g. the ‘sentence’ *aaa aa*. In the sentence-as-string there would be 3 occurrences of *aa*. In the sentence as sequence of strings, where the *substrings* are the strings occurring in the sequence, just one.

We see three shortcomings to Wetzel’s analysis. First, she fails to note that the notion of occurrence cuts across the type-token distinction: she places occurrences squarely at the level of types. However, it seems to us that the notion of occurrence also makes sense at the token level. We grant that it is a subtle matter to separate occurrences at the token level from tokens. The reason is that

---

a (1)-token. Conversely, given a (2)-token  $\theta$  of type  $\Theta$ , we can model the corresponding (1)-token as  $\langle \theta, \Theta \rangle$ . The weakness of the argument is that it is not clear that the class of contours that realise some property will automatically be a kind. Similarly, will the pairs modelling (1)-tokens go proxi for the members of a kind? At most the argument shows that to distinguish between (1) and (2) we need additional insight in what it is to be an object or a kind of objects.

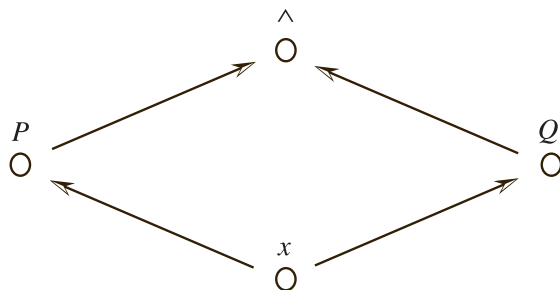
a separating argument always rest on some choices concerning the individuation of tokens. For example, in our leftmost picture above, we have —*under one possible analysis of what a token is*— one token of ‘S’ carrying two distinct occurrences.<sup>14</sup>

We could model token occurrences —Wetzel style— as follows. Suppose we have a token text / string  $\alpha$  of type  $A$ . Let  $a$  be a type occurrence  $\langle n, \tau, A \rangle$ . Then we can take  $\langle \alpha, a \rangle$  to be a token occurrence of a string of type  $\tau$  in  $A$ .

Secondly, Wetzel’s solution will individuate occurrences in a rigid static way. Consider an occurrence  $a$  of  $\rho$  in  $\sigma$  and an occurrence  $b$  of  $\sigma$  in  $\tau$ . Then, clearly, there is an occurrence  $c$  of  $\rho$  in  $\tau$ , given by  $a$  and  $b$ . Now think of  $\sigma$  and  $\tau$  as (types of) stages of an expanding text. This text is to be thought of as given as a concrete process. It seems natural to want to speak about an occurrence of  $\rho$  in the growing text corresponding to  $a$  and  $c$ . Things get even more complex when the text is not just growing but also being modified (as was the text of this paper).

The problem of a full modelling of the notion of occurrence in a changing world is beyond the scope of this paper. We just want to point out to the reader that in the context of term rewriting some ideas on how to do this have been developed employing *labelling*. See e.g. (Terese, to appear).

Thirdly, as indicated in our discussion above about the sentence-as-string vs. the sentence-as-sequence-of-strings, what counts as occurrence will be dependent on the ‘level’ of analysis we are considering. E.g. when looking at a parse tree of a sentence, we may well wish to identify occurrences with the nodes of the tree. But now consider the formula in predicate logic  $(P(x) \wedge Q(x))$ . This formula has two occurrences of the variable  $x$ . An attractive way to parse this formula is as a sharing graph:



The graph only contains one occurrence of  $x$ .

## 5 Words in Time

Words have a temporal dimension in two ways. First, a token of a word can only be a token of that word by being in an appropriate context. A major in-

<sup>14</sup>A theme that we ignored is preservation under the copying relation. Consider, for example, the token-text *Macavity*, *Macavity*, there’s no one like *Macavity* as written by Wetzel when writing her article. There is a sense in which the text reproduced in my copy of the *Journal of Philosophical Logic* is the same text as the text Wetzel wrote. This sense is a stronger sense than just identity of type. So, maybe we should say, that e.g. the first occurrence of *Macavity* in the text as written by Wetzel is the same occurrence as the first occurrence in the text in my copy of *JPL*?

gradient of such a context is constituted by the fact that the use of the word of which the token is an embodiment, stands in such-and-such a connection with earlier uses of the word. Thus, a word is essentially historical. A word without a story isn't a word.

Secondly, a word may actually change over time while still remaining *that word*. In fact, every one of the 'components' of a word, like sound, spelling, gender, meaning can be subject to change.

Of course, some changes destroy sameness. Consider the story of loanwords. The Dutch word bolwerk indicates a fortification and the pentagonal architecture around it. It was adopted as a loanword into various languages, including English (where it is written as *bulwark*) and French (as *boulevard*). The French word subsequently gradually changed its meaning to a 'former bulwark transformed into a walking path', and as such was adopted back into Dutch as a French loanword (and changed once again until it simply meant 'broad street'). But the two Dutch words bolwerk and boulevard are not the same word, despite their historical relatedness. The same holds for the Dutch (Belgian) word mannekijn, used for the wooden dolls used by tailors, which was adopted into French and back again to become mannequin.

So, clearly, along some paths identity is broken. However, it is equally clear that e.g. a spelling reform does not change the stock of Dutch words even if a number of them are suddenly written in a different way. Pannenkoek is still pannekoek! It even does not seem not plausible to hold that a word always must have one fixed meaning. A well-known example is that the English word nice has radically changed its meaning (and spelling) over time, as can be found in the Oxford English Dictionary. It is derived from the word *nescius* (ignorant), and used to be a negative word: in the 14th and 15th century it meant *foolish or stupid* (*they seiden he was a fool... and that they sien neuere so nise a man*, ca. 1450). By 1529, it got to mean *wanton*, or *loose-mannered*. It then shifted to *tender, reserved*, and *requiring great precision*, which lead in the 18th to *dainty or appetising*, esp. of a cup of tea (*we sent her up three or four plates of the nicest things that were at table*, 1766). The meaning it now has (agreeable) is incomparable with its original meaning, and its spelling changed from *nise* to *nice*. The general tendency is to still call this the same word. Clearly, there is an uninterrupted historical chain between the 15th century nise and the current-day nice. But as shown by the bolwerk example, this is not enough for sameness. There is no clear definition of when the chain of sameness is broken, and also the question whether the 15th century nise and the current-day nice should count as the same word is open for debate. This once more illustrates the illusive nature of words.

## 6 So what is a Word really?

A word is nothing like an equivalence class of equiform objects or anything like that. Such a view of wordhood tries to make words into neutral objects separated from the ways in which words are a part of practice. Words exist in uses of words. A use of a word presupposes a historical connection to earlier

uses of the same word.<sup>15</sup>

A word is not an abstract object like a number or a form. Words have a beginning in time, words are spread around, words change, words disappear. They are temporal objects.

It is useful to associate to a word a set of vectors. The co-ordinates of these vectors are projections of the word. the co-ordinates are things like: word-classes, pronunciation, spelling, hyphenation, gender, denotation . . . Note that a word relates in a quite different way to e.g. its pronunciation and its meaning. The pronunciation has to do with the physical realisation of the word. The meaning with why we use it at all. So the vector representation should not be misunderstood as the idea that all co-ordinates are on a par. The sets are understood disjunctively: we need to model variations in spelling, meaning, etc. The set of vectors *is* not the word, but it summarises major information about the word. We get a notion like *wordform* by restricting the vectors to certain dimensions.

The word word is really ambiguous. We can consider is and are as different words. We can also consider them as the same word. The notion of *lexeme* captures much of the word in this second sense.

We end with a metaphor —admittedly misleading as all metaphors are. Compare a word to a bank account. Bank accounts are only bank accounts because we behave as we behave. A bank account has a history. A bank account is neither equal to the amount of money on it, nor to the account number. In fact both may be subject to change. (Account numbers may change in a reorganisation of the numbering system, e.g. after a merger of two banks. This is a bit like spelling reform.) Still at any given moment the pair account number, amount gives important information about the account. A bank account is moderately abstract: it is nowhere and everywhere. The pronunciation and spelling of a word can be compared with the account number: they are modes of access. The meaning of a word can be compared to the money on the account; the represent the value. A word is a kind of shared account we keep meanings on.

## References

- Aristotle. 1966. *Categories, and De Interpretatione*. Oxford: Clarendon Press.  
Translated with notes by J.L. Ackrill.
- Béjoint, Henri. 1994. *Tradition and Innovation in Modern English Dictionaries*.  
Oxford: Clarendon Press.
- Corstius, Hugo Brandt. Hier en Nu. *NRC Handelsblad, Cultureel Supplement*,  
February 13 year =:CS8.

---

<sup>15</sup>The notion of *use* should be taken here *cum grano salis*. A man is taking money from an ATM. He reads: *do you want to know your saldo?*. Clearly the man is reading words. In fact, the word *you* denotes the man. However, the machine is not uttering this question, neither are the programmers of the machine. The programmers just brought it about that similar questions would be asked anytime the machine was used. Everytime it's a different question however since addressee and question time vary. In such an example one would like to say: the words get used without there being an utterer/user.

- Frege, G. 1976. Der Gedanke. *Pages 30–53 of: Günther Patzig (ed.), Logische Untersuchungen.* Göttingen: Vandenhoeck & Ruprecht.
- Frege, G. 1988. *Die Grundlagen der Arithmetik.* Hamburg: Felix Meiner Verlag.
- Kaplan, David. 1990. Words. *Aristotelian Society, Supp.* 64:93–119.
- Katamba, Francis. 1994. *English Words.* London: Routledge.
- Lamport, Leslie. 1994. *LaTeX User's Guide and Reference Manual.* Second edn. Reading: Addison-Wesley.
- Lyons, John. 1995. *Linguistic Semantics: an Introduction.* Cambridge: Cambridge University Press.
- Manguel, Alberto. 1996. *A History of Reading.* New York: Penguin Putnam Inc.
- Mel'čuk, Igor A. 1993. *Cours de Morphologie Générale.* Vol. 1: Introduction et Première Partie: le Mot. Montréal: Les Presses de L'Université de Montréal.
- Messelaar, P.A. 1990. *La Confection du Dictionnaire Générale Bilingue.* Leuven: Peeters.
- Peirce, Charles S. 1906. Prolegomena to an Apology for Pragmaticism. *In: C. Hartshorne & P. Weiss (eds.), Collected papers of Charles Sanders Peirce.* Cambridge: Belknap Press of Harvard University Press. 1960.
- Quine, Willard Van Orman. 1960. *Word and Object.* Cambridge: MIT Press.
- Reichling, Anton Joannes Bernardus Nicolaas. 1989. *Het Woord: een studie omtrent de grondslag van taal en taalgebruik.* Ph.D. thesis, Rijks-Universiteit Utrecht, Nijmegen.
- Saussure, Ferdinand de. 1972. *Cours de Linguistique Générale.* Paris: Édition Payot.
- Sciullo, Anna-Maria di. 1988. *On the Definition of Word.* Cambridge: MIT Press.
- Tarski, Alfred. 1944. The semantic conception of truth and the foundations of semantics. *Philosophy and Phenomenological Research*, 4:341–375.
- Tarski, Alfred. 1994. The semantic conception of truth and the foundations of semantics. *Pages 536–570 of: R.M. Harnish (ed.), Basic Topics in the Philosophy of Language.* New York: Harvester Wheatsheaf.
- Terese (ed.). to appear. *Term Rewriting Systems.* Cambridge: Cambridge University Press. Chap. 8: Equivalence of Reductions.
- Wetzel, Linda. 1993. What are occurrences of expressions? *Journal of Philosophical Logic*, 22:215–219.
- Wittgenstein, Ludwig. 1971. *Philosophische Untersuchungen.* Frankfurt am Main: Suhrkamp Taschenbuch Verlag.