The Whorfian hypothesis and numerical cognition: is 'twenty-four' processed in the same way as 'four-and-twenty'?

Marc Brysbaert[a,b,*], Wim Fias[a], Marie-Pascale Noël[c]

[a]*Laboratory of Experimental Psychology, Katholieke Universiteit Leuven, Leuven, Belgium*
[b]*Department of Experimental Psychology, University of Ghent, Ghent, Belgium*
[c]*Department of Cognitive Neuropsychology, Université Catholique de Louvain, Louvain, Belgium*

Abstract

Recent theoretical developments have redefined a Whorfian effect as a processing difference due to the language of the individual, and no longer as a marker for or against linguistic determinism. Within this framework, Whorfian effects can be used to investigate whether a particular part of the cognitive system is penetrable by language processes or forms an encapsulated module, provided the experimenter ensures that the target language difference is not caused by peripheral input or output processes. In this article, we examine the possibility of a Whorfian effect in numerical cognition by making use of the fact that in the Dutch number naming system the order of tens and units is reversed (i.e. 24 is read 'four-and-twenty'). In a first experiment, we asked native French- and Dutch-speaking students to name the solution of addition problems with a two-digit and a single-digit operand (e.g. $20 + 4 = ?$, $24 + 1 = ?$). The order of the operands was manipulated ($20 + 4$ vs. $4 + 20$) as well as the presentation modality (Arabic vs. verbal). Three language differences emerged from this study. Experiment 2, however, showed that these differences were all due to input or output processes rather than differences in the addition operation (i.e. the differences between Dutch and French disappeared when subjects were asked to type the answer rather than pronounce it). On the basis of these findings, we question the idea that mathematical operations are based on verbal processes. © 1998 Elsevier Science B.V.

Keywords: Whorfian hypothesis; Numerical cognition; Mathematical operations

* Corresponding author. Department of Experimental Psychology, University of Ghent, Henri Dunant-laan 2, B-9000 Ghent, Belgium. Tel.: +32 9 2646436; fax: +32 9 2646496; e-mail: Marc.Brysbaert@ru-g.ac.be

1. Introduction

At the beginning of this century anthropologists became aware that their views of other cultures had been ethnocentrically biased. This awareness, usually credited to Boas, gradually resulted in the conviction that the language people use may have a major impact on the way they perceive reality, the so-called Sapir–Whorf hypothesis (Sapir, 1949; Whorf, 1956). Although the idea was originally meant as a warning not to treat language as a set of labels on a pre-existing, non-cultural (or 'objective') world, by the middle of the 1950s, a scholarly folklore had grown around Sapir and Whorf that hardened 'linguistic relativity' into 'linguistic determinism' (Hill and Mannheim, 1992, p. 385). According to this view, all human thoughts are shaped by the language that is used. As such, the Whorfian hypothesis was rapidly disconfirmed, so that '... in most circles of experimental psychology it is impossible to mention Whorf's thesis without quick acknowledgement of its empirical disconfirmation'. (Hardin and Banaji, 1993, p. 279).

1.1. Linguistic determinism dismissed

The early, devastating, evidence against the idea of linguistic determinism came from two lines of research. Because these have been documented elsewhere (see in particular Gerrig and Banaji, 1994; and Hardin and Banaji, 1993), only a summary of them will be given here. The first line of research concerned colour memory. As Whorf had suggested that language users 'dissect nature along the lines laid down by (their) native languages' (Whorf, 1956, p. 213), colour seemed one of the most promising candidates to test the idea. The colour spectrum was thought to provide a continuous gradation of stimuli to which labels may be arbitrarily assigned. Because colour labels were known to vary across cultures, linguistic relativity would be supported if corresponding differences in colour recognition could be identified across linguistic communities. Although the first studies were encouraging, the effects were small and shortly afterwards overruled by two major publications (Brown, 1976). In the first publication, Berlin and Kay (1969) claimed that all colour terms of all languages could be reduced to a total of eleven colours, and furthermore, that colour terms emerged across linguistic communities according to a five-level hierarchy: (1) black, white, (2) red, (3) yellow, green, blue, (4) brown, and (5) purple, pink, orange, grey. Thus, if a particular language had just two basic colour terms, the terms would correspond to 'black' and 'white'. If a language included a third term, it would be 'red', and so on. As the hierarchy seemed in line with the physiological aspects of colour, this finding was taken as evidence that human physiology determined the development of linguistic terms and placed limits on interlinguistic variation in colour perception.

Even more damaging for the hypothesis, Heider (Heider, 1972; Heider and Olivier, 1972) showed that the Dani of New Guinea, who have only two basic colour terms, were nevertheless better able to recognise previously presented focal than non-focal colour chips, similar to the whites of North America. From this evidence, Heider argued that perceptual salience, not language, caused differences in memory,

thereby compromising even the hypothesis of linguistic relativity (Hardin and Banaji, 1993, p. 283).

The second line of research that disconfirmed the idea of linguistic determinism came from counterfactual reasoning (e.g. Gerrig and Banaji, 1994, pp. 239–242). Given that Chinese provides no straightforward means for marking counterfactuals (as in 'If he were Sara's teacher, Sara would do better at school'), Bloom (1981) hypothesised that Chinese speakers would be less able than English speakers to recognise counterfactual arguments. He presented English and Chinese subjects stories that contained counterfactual implications. For instance, one story told of a European philosopher who would have been able to contribute to philosophy if he had been able to read Chinese. The two sets of subjects were afterwards asked whether the philosopher had actually made the contributions alluded to in the text. As expected on the basis of linguistic determinism, English subjects vastly overperformed their Chinese colleagues (98% correct responses against 6% correct responses). However, Au (1983) shortly afterwards showed that Bloom's Chinese texts were not good translations and provided a different meaning than the one he had intended. When Au corrected Bloom's Chinese, all evidence for an influence of language on thought disappeared: English and Chinese speakers both performed at near-perfect rates (see also Liu, 1985).

1.2. The resurgence of the Whorfian idea

Recently, a number of attempts have been published to rehabilitate the Whorfian hypothesis. In general, these are based on three arguments: (1) some of the early counter-evidence needs to be reconsidered, (2) linguistic determinism is too strong a variant of the Whorfian hypothesis, and (3) there are more refined ways to look at Whorfian effects than to verify whether cultures can or cannot perform a certain task.

As regards the first argument, some researchers have examined the evidence against the Whorfian hypothesis and argued that it is not as strong as it first seemed to be. For instance, a closer look at the evidence provided by Berlin and Kay (1969) shows that the empirical part is rather thin (for a revealing review, see Saunders and van Brakel, 1997). For a start, it is quite amazing that the 11 basic colour terms coincide with the most popular American–English colour terms in Thorndike's 'Teacher's Handbook' (ethnocentricity?). Second, of the 20 languages for which Berlin and Key gathered data, 19 were represented by only one bilingual speaker, who lived in the neighbourhood of the university. Finally, there is concern that a considerable degree of subtlety of the original colour names was disregarded to make them fit within the framework.

Heider's research was also criticised: the array of colour chips she had used to test both the Dani and English speakers appeared to be biased in a way that made the focal colours a priori more salient than the non-focal colours (they were the ones with the highest saturation). Lucy and Shweder (1979) constructed a new test array that was not subject to this bias and failed to replicate Heider's original findings. They demonstrated, in fact, that what mattered most for accurate recognition mem-

ory was not focality, but the availability of a 'referentially precise basic colour description' (p. 159). Further demonstrations along these lines were provided by Kay and Kempton (1984); see Gerrig and Banaji (1994, pp. 238–239) for a summary.

The second argument contributing to the resurgence of the Whorfian hypothesis is more theoretical. It basically says that the rejection of linguistic determinism is not in contradiction with the possibility that language may exert a significant influence on particular parts of (non-verbal) cognition. To state it sharply, there is a difference between testing the hypothesis 'language shapes all cognitive processing' and the hypothesis 'language has no effect on cognitive processing'. Whereas one demonstration of a lack of Whorfian effect suffices to discard the former hypothesis, the finding of a single Whorfian effect is already enough to reject the latter. So, the real research question is whether Whorfian effects exist.

Finally, a third argument has been introduced by Hunt and Agnoli (1991). According to them, only the strongest version of the Whorfian hypothesis predicts that a thought expressible in one language must not be expressible in another. A weaker, more plausible, form is 'that language differentially favours some thought processes over others, to the point that a thought that is easily expressed in one language might virtually never be developed by speakers of another language.' (p. 378). So, if an intralanguage effect of 50 ms is considered as a robust effect, it suffices to find a cross-linguistic difference of the same magnitude to show a Whorfian effect.

Hunt and Agnoli (1991) frame their suggestion within a view of cognition that is based on three levels at which cognition can be studied (cf. Pylyshyn, 1984). The lowest level is concerned with the physiological mechanisms underlying thought. These are presumably cultural universals. The highest level is the representational level and is concerned with meaning. In-between are the mechanics of thought, which largely rely on language. They convert the visual or auditory input code into a symbol meaning structure that represents a non-linguistic reality. These mechanisms have a cost of computation, which partly depends on the language. As people consider the costs of computation when they reason about a topic, this will make language influence cognition. For instance, Hunt and Agnoli note that English contains many more polysemous words than Italian, and wonder what the cognitive consequences of this could be. They also mention a study of Hoffman et al. (1986) which indicated that spontaneous labelling by language users influences their memory for social or ill-structured perceptual events. Hoffman et al. (1986) showed descriptions of persons to bilingual English–Chinese speakers. The descriptions were chosen in such a way that the character traits were part of different stereotypes for the two languages. So, the stereotype one could infer depended on the language used to describe the person. Subsequently, subjects were asked whether particular behaviours not given in the original description were likely to be characteristic of the target person. English speakers extrapolated traits associated with the stereotype of their language. Chinese speakers addressed in Chinese used Chinese stereotypes, but when addressed in English they used English stereotypes.

To Hunt and Agnoli's (1991) basic claim, two more constraints should be added (or made explicit) in our view. First, it is not enough to show a significant difference in the processing (time) between two languages, it is also necessary to demonstrate that this difference is not due to peripheral input or output processes. For instance, it has been shown that the digit memory span is shorter for Welsh than for English (Ellis and Hennely, 1980; Hoosain, 1986; Naveh-Benjamin and Ayres, 1986). However, can we really consider this as a Whorfian effect, if it is known that the memory span depends on the length of the stimuli (Baddeley et al., 1975), and that digit names are longer in Welsh than in English? (but see Hardin and Banaji, 1993).

A second constraint we want to introduce, is that the effect should go beyond mere language processing. For instance, it has been claimed that reading processes may differ as a function of the script (e.g. logographic vs. alphabetical) and the transparency of the letter–sound relations. However, one would hesitate to call these differences Whorfian effects, as they concern the mechanisms of language processing itself (but see Hoosain, 1986). In our view, better examples of Whorfian effects are (i) the finding that subjects who do not read an alphabetical script, experience difficulties in a number of tasks ranging from phoneme manipulation (Morais et al., 1986; Morais and Kolinsky, 1994; Read et al., 1986) to parts detection within a complex visual figure (Kolinsky et al., 1990), and (ii) the finding that sound discrimination is already influenced by the surrounding adult language in infants of less than a year (e.g. Werker and Tees, 1984; Mehler et al., 1988). These findings look very much like real Whorfian effects.

In summary, if one agrees with Hunt and Agnoli's (1991) formulation of the Whorfian hypothesis (i.e. that language can affect the cost of some cognitive processes), the search for Whorfian effects gets another meaning. Rather than a quest for or against linguistic determinism, it becomes an investigation to what extent a particular part of human cognition is penetrable by the language system.

In the remainder of the text, we will look at Whorfian effects in number processing. Possible language influences are of great interest here, because a lot of theorising is centred around the question of whether numerical operations are performed on a single (non-verbal) numerical input format or not. Virtually all literates are familiar with at least two notations: Arabic and verbal (written and auditory). In addition, small numbers can be represented in an analogue way by showing figures with a different area or by varying the number of dots on a display (e.g. on a dice). So, one of the major questions about numerical cognition is how the different notations (one of which is verbal) are recognised and handled in various operations (for recent reviews see Brysbaert, 1995; Blankenberger and Vorberg, 1997; Noël et al., 1997; Noël and Seron, 1997).

After scrutinising the available evidence for Whorfian effects in number processing, we will consider the major current theories and examine whether these do or do not predict language differences. Then, two new experiments will be presented that investigate the possibility of a Whorfian effect due to number naming in French and Dutch.

1.3. Whorfian effects in numerical cognition

In several texts, language differences in numerical cognition have been given as examples of potential Whorfian effects. For a start, Hunt and Agnoli (1991) pointed to the fact that a lot of our number schemes are not present in other languages. For instance, if we say that there are 37 pairs of shoes for 49 men, most English speakers will understand that some men will have to go without shoes. This situation may be much more difficult to describe in a non-literate society where the language may have number terms only for 'one-two-many' (Greenberg, 1978). Another example given by Hunt and Agnoli (taken from Saxe and Posner, 1983) has to do with languages that utilise body analogies to define numbers. Children who learn these languages, experience difficulties in counting that are related to symmetries in the body part used to represent numbers, and cannot perform more complicated concepts and operations of mathematics, such as division or the distinction between rational and irrational numbers.

Unfortunately, it is not always clear in these examples what part of the cultural differences is due to language and what part to the fact that other societies may not be so interested in the mathematics we find important. In the latter case, the language difference may come very close to the distinction between those who, for example, understand matrix algebra and those who do not. However, Hunt and Agnoli also pointed to subtler but real differences in arithmetic capabilities associated with linguistic differences in fully literate societies.

The Asian languages have a much simpler number naming system than most Western languages. The latter usually have irregular names for teens (eleven, twelve, thirteen,...) and decades (twenty, thirty,...). In contrast, the former languages have the teens named as other two-digit numbers ('ten one', 'ten two',...) and the decades by multiplying the ten's name ('two ten', 'three ten',...). There is now repeated evidence that this notation helps Asian children to understand the number system more quickly and more profoundly (e.g. Miller, 1996; Miura et al., 1993). For instance, when young Asian children are asked to represent numbers with two types of blocks that stand for tens and units, they are likely to use both types of blocks, whereas Western children of the same age have a higher tendency to represent the numbers by an equivalent amount of unit blocks. There have been claims that this could be one of the origins of the superior mathematical achievements of Asian cultures compared with Western cultures, although this view is not shared by everybody (e.g. Geary (1996) points to large differences in schooling and cultural valuation; and Towse and Saxton (1997) point to methodological problems with Miura's task).

An analogous finding was reported by Seron and Fayol (1994) who compared French-speaking children from Belgium and France. The verbal number system in French-speaking Belgium (and Switzerland) is simpler than the one in France. This is because the Belgians use the names 'septante' for seventy and 'nonante' for ninety, whereas the French use the names 'soixante-dix' (sixty-ten) and 'quatre-vingt-dix' (four times twenty-ten). In addition, all number names from 70 to 79 in France are constructed by combining 'sixty' with a teens name (soixante-et-onze,

soixante-douze,...), and all number names from 90 to 99 by combining 'eighty' with a teens name (quatre-vingt-onze,...). Seron and Fayol reported that second-grade children in France made more errors in Arabic number production than their Belgian counterparts (especially for the seventy and the ninety decades) and that not all differences could be explained by production difficulties (i.e. problems in the actual writing of Arabic numerals). That is, part of the difference could be defined as a genuine number comprehension deficit. This observation agrees with previous observations by Deloche and Seron (1982), who reported similar differences in performance for Belgian and French adult aphasics.

Still in relation to the issue of teens' names, Miller and Zhu (1991) observed that English undergraduates experienced difficulty reversing two-digit numbers ending in one (e.g. seeing '71' and saying 'seventeen'). The same difficulty was present for Chinese–English bilinguals when performing the task in English, but not in Chinese. A problem with Miller and Zhu's study, however, is that the effect may well have been situated in the language production stage of the task. That is, subjects may have experienced difficulties not because the names of the teens were morphologically anomalous, but because there was interference from the stimulus name in the speech output stage (i.e. from seventy-one on seventeen; see especially experiment 2 of Miller and Zhu (1991) and their own acknowledgement of this interpretation problem).

All in all, language differences in numerical performance have been reported on several occasions. Unfortunately, only a few of these studies have related their findings to theories of numerical cognition and tried to get some more information about the processing stage at which the language effect could be situated. Furthermore, most of these effects were limited to situations in which subjects were acquiring the skills under investigation. This is not really strong evidence for the Whorfian hypothesis, as it might imply that language is only needed for the creation of an autonomous numerical system (a view that probably would not be opposed by the proponents of theories that are based on a non-verbal code, see below). What would constitute real counter-evidence for non-verbal theories of numerical cognition, is the demonstration of a Whorfian effect for a well-known, overlearned operation such as, for instance, the realisation of simple mathematical operations.

In Sections 1.4 and 1.5, we will consider the different theories of number processing and calculation and examine whether or not they predict a Whorfian effect for simple mathematical operations.

1.4. Theories of numerical cognition that do not accept Whorfian effects

Two theories of number processing do not accept the possibility that the language of the individual influences his numerical performance. The first has been proposed by McCloskey and colleagues (e.g. McCloskey, 1992; McCloskey and Macaruso, 1995). It assumes that all numerical input, independent of its notation, is converted to an amodal abstract semantic representation before further processing is possible. The semantic system contains representations that specify in abstract form the basic quantities of a number and their associated powers of 10. So, a number like 4020 is

represented as four times 10 to the third power (or thousand) and two times 10 to the first power. As soon as a numeral has been converted into its abstract representation, all information about the surface form is lost. The semantic system is the starting point for arithmetics and all modality-specific output mechanisms.

The second theory of number processing that is not likely to incorporate language effects, is Gallistel and Gelman's (1992) preverbal counting model, which states that numerical cognition depends on analogue representations that are innate and already present in non-human species such as pigeons and rats. These representations have emerged throughout evolution because for many species counting is essential for survival. Humans use their innate representations when they acquire verbal competence with numbers. In particular, Gallistel and Gelman (1992) argue that when children learn to count, they learn a bi-directional mapping between the preverbal magnitudes that represent numerosity and the number words. The preverbal magnitudes are also used to do basic calculations such as single-digit addition and multiplication (it is not clear from Gallistel and Gelman's writings how, for example, two-digit additions are handled).

Finally, there are also fact retrieval models of basic arithmetics that do not specify the representations on which the retrieval network is grounded (e.g. Ashcraft, 1992), and therefore do not enable precise predictions.

1.5. Theories of numerical cognition that tolerate Whorfian effects

Obviously, the theories that are most favourable to the existence of Whorfian effects in number representation and number operations, are theories which claim that numerical cognition depends on language. Such a view has, for instance, been defended by Hurford (1987), who wrote that: 'the number faculty largely emerges through the interaction of central features of the language faculty with other cognitive capacities relating to the recognition and manipulation of concrete objects and collections.... It is therefore not necessary to postulate an autonomous 'faculty of number' as a separate module of mind.' (p. 3). However, this author did not propose a theory or a model of number processing and calculation per se.

Campbell and Clark's (Campbell and Clark, 1988; Campbell, 1994) multiple code model also predicts the appearance of Whorfian effects in numerical cognition. According to these authors, number processing (including mathematical operations) operates on a complex of multiple representations some of which are verbal. Consequently, effects of language differences can be expected. However, in a subsequent article, Campbell (1994) claimed that the code on which mathematical operations are performed depends on the input, so that according to this latest version of the multiple code model language differences should be more prominent for verbal numerals as input than for Arabic numerals.

Other theories claim that not all numerical cognitions depend on language, but that some aspects do. For instance, in his triple code model, Dehaene (1992) postulates three interconnected number systems: An amodal analogue magnitude system, a visual Arabic system, and an auditory verbal system. Each of these systems is

the starting point for different numerical activities. The analogue magnitude system is used for approximate calculation and magnitude comparison, the visual Arabic number form is addressed for multi-digit operations and parity judgement, and the auditory verbal word frame is used for counting and basic number fact retrieval (i.e. single-digit additions and multiplications). The idea that basic fact retrieval depends on a verbal phonological storage is motivated by several observations (summarised in the review article of McCloskey and Macaruso, 1995, p. 360). First, children learning arithmetic facts typically engage in substantial amounts of overt and covert oral rehearsal, perhaps leading to the creation of phonological memory representations for these facts. Second, children and adults often report that they say simple arithmetic problems to themselves when trying to remember the answers. Third, multilingual individuals appear to perform calculations in the language they spoke when learning arithmetic (Kolers, 1968; Shanon, 1984; but see Marsh and Maki, 1976; and McClain and Shih Huang, 1982 for an absence of language effect on bilinguals' calculation). So, Dehaene's (1992) model might predict a Whorfian effect in simple mathematical operations (which depend on verbal phonological representations), but would not expect such an effect in number comparisons (which depend on the amodal, analogue magnitude system).

Finally, Noël and Seron (1997) reported that number manipulations could be based on intermediate representations that are a function of the input's lexico-syntactic structure. For instance, they argued that the Roman numeral 'VI' activates the representation of the sum of the quantities {5} and {1} whereas the word numeral 'six' only activates the representation of the quantity {6}. In line with this reasoning, they showed (experiments 1–3) that the verification of 'VI = 5 + 1' is much faster than the verification of 'VI = 3 + 3', and that this difference is seriously reduced for the trials 'six = 5 + 1' and 'six = 3 + 3'. Similarly, they proposed that 'douze cents' (twelve hundred) and 'mille deux cents' (one thousand two hundred) activate different intermediate representations: 'douze cents' would activate the representation of $\{12\} \times \{100\}$ whereas 'mille deux cents' would activate the representation of $\{1000\} + \{\{2\} \times \{100\}\}$. They argued that these intermediate representations influence some aspects of number processing and calculation. In support of their claim, Noël and Seron showed (experiment 4) that indicating which member of a verbal number pair is the largest depends on the lexico-syntactic structure of the pair, with subjects being much faster to respond when both numerals have the same lexico-syntactic structure (e.g. twelve hundred/thirteen hundred) than when they have a different structure (e.g. twelve hundred/one thousand three hundred). Furthermore, effects of the lexico-syntactic structure of the numerals were also obtained in calculation tasks (experiments 5 and 6). For instance, Noël and Seron showed that the lexico-syntactic structure of the addend has a significant influence on the verbal structure used to express the answer. Thus, to the problem 'twelve hundred + three hundred', people tended to answer more frequently 'fifteen hundred' than to the comparable problem 'one thousand two hundred + three hundred'. On the basis of these results, Noël and Seron concluded that code-dependent intermediate representations can be used in mathematical operations, and hence that language influences may be expected in some specific situations.

1.6. Brief presentation of the experiments

Two experiments are reported in this article. They were designed to find out whether we can elicit a Whorfian effect in numerical cognition by using a manipulation that is very similar to the one proposed by Noël and Seron (1997), (see Section 1.5) but with stimulus materials that are much closer to everyday practice and, hence, more likely to be fully acquired (see the problem of the acquisition phase discussed above). More precisely, we took advantage of a difference in number naming between two different languages (French and Dutch) that are spoken by people who live in nearby regions of the same country and therefore are likely to share the same cultural and educational background. We made use of the fact that in Dutch (unlike French) two-digit numbers are named in reversed order (e.g. 24 is pronounced 'four-and-twenty' instead of 'twenty-four')[1]. We looked at what is probably the strongest effect one can expect from this practice, namely the effect the number name has on the addition of its own constituents. If numerical addition is based on language processes, one could expect that in Dutch the problem 4 + 20 is easier to solve than the problem 20 + 4, whereas in French the reverse should happen. In experiment 1, the realisation of problems such as 20 + 4 and 4 + 20 in both the Arabic and the written verbal code will be compared for French- and Dutch-speaking subjects. In experiment 2, we will further elaborate the findings by looking at an output modality (typing instead of naming) that requires the Dutch-speaking subjects to respond like the French-speaking (i.e. by starting with the value of the ten).

It may be noted that by looking at the performance of people with a different native language, we circumvent one of the main problems in research dealing with the relationship between input format and mathematical cognition (Noël et al., 1997; Noël and Fias, 1998). Because most of the research is based on designs with repeated measures, it is often difficult to determine whether the input effects are due to differences in the semantic numerical system or to differences in the encoding and production stage of the different types of materials. For instance, it has repeatedly been shown that bilinguals need more time to calculate in their second language than in their first language. However, this effect may be caused by a difficulty in translating a numeral of the second language to an abstract representation or to the corresponding numeral of the first language, as well as to differences in the semantic system(s) underlying the mathematical operation (Noël and Fias, 1998). This is particularly important for the evaluation of a model like Dehaene's, which claims that basic number fact retrieval relies on auditory verbal representations of the first language (Dehaene, 1992, p. 33).

A difference in processing time for the problems 20 + 4 and 4 + 20 as a function of the subject's language would be at odds with models that assume arithmetical facts retrieval to be realised on a non-verbal, abstract (McCloskey, 1992) or analogue (Gallistel and Gelman, 1992) representation. Such a language difference, how-

---

[1]The reversal of tens and units in the names of two-digit numbers also occurs in German, and was present in some British dialects up to the last century. So, in Jane Austen's *Pride and Prejudice* (1813, Volume 1, Chapter 9), Mrs Bennet proudly announces: '... I know we dine with four and twenty families!'.

ever, would be in line with the encoding complex model of Campbell and Clark (1988), the triple-code model of Dehaene (1992), and the intermediate representation hypothesis of Noël and Seron (1997). In addition, by using an Arabic as well as a verbal presentation format, we can find out whether both presentation formats are converted to the same code before the arithmetic operation takes place (as predicted by McCloskey, Gallistel and Gelman, and Dehaene), or whether they elicit different processes (as assumed by Campbell, 1994; and Noël and Seron, 1997). Campbell (1994) would predict a larger effect of language for verbal presentations than for Arabic presentations because he assumes that the code on which the retrieval of arithmetical facts occurs depends on the input format and, hence, that language effects are more likely for the verbal than for the Arabic input. Similarly, Noël and Seron's hypothesis predicts a larger effect for the verbal than for the Arabic input, because only for the verbal code is there a difference in intermediate representation ($\{4\} + \{20\}$ instead of $\{20\} + \{4\}$).

2. Experiment 1

In this experiment, we compared naming latencies of French- and Dutch-speaking individuals when they had to pronounce as fast as possible the solution of a simple addition that consisted of a two-digit number and a one-digit number (e.g. $20 + 4 = ?$, $3 + 45 = ?$, $67 + 8 = ?$). The two-digit numbers were numbers between 20 and 99, the single-digit numbers ranged from 1 to 9. The order of numbers was arbitrary (at least from the subject's point of view); that is, on some trials the two-digit number was first, on other trials it was last. All possible sorts of combinations of single- and double-digit numbers were used, also those resulting in a carry-over effect (e.g. $9 + 74 = ?$) and those that crossed the boundary of 100 (e.g. $95 + 6 = ?$). Finally, half of the stimuli were presented in the Arabic format, half in the verbal format. Format was blocked.

2.1. Method

2.1.1. Subjects
Twenty-four Belgian native French-speaking students from the Université Catholique de Louvain and 24 Belgian native Dutch-speaking students from the Katholieke Universiteit Leuven participated on a voluntary basis. The Dutch-speaking University of Leuven and the French-speaking University of Louvain are situated in independent campuses some 30 km away from one another. Before 1972, they were part of the unitary University of Leuven/Louvain. The French-speaking subjects used the Belgian number naming system (i.e. with septante and nonante; see Section 1.3).

2.1.2. Stimuli
Stimuli were addition problems with one addend between 1 and 9 and the other addend larger than or equal to 20. Two factors were manipulated: order (tens + units

or units + tens) and input modality (verbal modality or Arabic modality). One third of the 216 stimuli were experimental items of the form ten (i.e. a two-digit number divisible by 10 (e.g. 20, 30, etc..)) + unit, or the reversed order. Four stimulus lists were created according to a Latin square. Each list contained all 72 possible combinations of a ten and a unit (i.e. 20 + 1, 20 + 2,..., 30 + 1, 30 + 2,..., 90 + 8, 90 + 9), but each item was assigned to a different order by modality condition (e.g. 20 + 1 was presented as '20 + 1' in list 1, as '1 + 20' in list 2, as 'un + vingt' or 'een + twintig' in list 3, and as 'vingt + un' or 'twintig + een' in list 4). In this way, each subject saw only one version of each combination, and all combinations were seen by six subjects per language group.

The rest of the items (144) were filler items of the form ten-unit (i.e. a two-digit number not divisible by 10: from 21 to 99 excluding 30, 40,..) + unit (or reversed order). These items were generated at random at the beginning of the session, so that the fillers were different for each subject.

2.1.3. Equipment

Stimuli were presented on a monochrome screen in default MS-DOS text mode. Presentation was controlled by a PC-compatible computer (286 processor) that had a voice key connected to the game port. Reaction latencies were measured to the nearest ms using the software of Bovens and Brysbaert (1990).

2.1.4. Procedure

The presentation of an item consisted of the following sequence of events. First, a fixation line was presented on a blank screen for 500 ms; then, the item was shown with the left operand, the operator, and the right operand centred around the middle of the screen. In order to avoid the Arabic problems being seen in single glimpses, they were sufficiently separated (i.e. the left operand was presented on position 16 and the right operand on position 64 of the 80×25 character space). There was also sufficient space between the verbal numerals and the operator (this space depended on the length of the operands which could vary between five letter positions as in 'vingt' and 19 as in 'quatre-vingt-quatre'; the shortest Dutch numerals contained six letter positions, the longest 17). The problem stayed on the screen until the voice key was triggered (with a maximum of 10 s). At this moment, the problem disappeared and the experimenter typed in the response given by the subject, after which the screen was erased. The experimenter also noted whether time registration had been successful. The experiment consisted of two sessions: one in the verbal and one in the Arabic modality, counterbalanced across subjects. Each session included 20 training trials of the type unit–unit and 108 test trials. Between sessions there was a short break.

2.2. Results

In the remainder of the text, the following notations will be used: T + U refers to problems involving a ten and a unit (e.g. 20 + 4), TU + Unc refers to problems involving a ten-unit and a unit without carry-over (e.g. 21 + 4), TU + Uc refers to

problems with a ten-unit and a unit, and a carry-over operation (e.g. 22 + 9); finally T&U refers to the presentation order largest number (involving a ten) first and smallest number second (e.g. 21 + 4), whereas U&T refers to the reverse condition (e.g. 4 + 21). For the TU + Uc type of problems, only the solutions smaller than 100 were taken into account, because the difference in naming between French and Dutch disappears to a large extent for numbers of three digits (i.e. the names in both languages begin with the hundred).

Data were first analysed with a $2 \times 3 \times 2 \times 2$ analysis of variance that included the variables: language group (French, Dutch), problem type (T + U, TU + Unc, TU + Uc), format (Arabic, verbal), and order of presentation (T&U, U&T). For the percentage of errors (see Table 1), this yielded significant main effects of problem type (T + U = 4.1%, TU + Unc = 7.7%, TU + Uc = 14.2%; $F(2,92) = 79.5$, MSe = 62.8, $P < 0.01$; post hoc Newman–Keuls comparisons showed that all three percentages differ significantly from one another), format (Arabic = 7.0%, verbal = 10.3%; $F(1,46) = 19.5$, MSe = 78.5, $P < 0.01$), and order of presentation (T&U = 9.4%, U&T = 7.9%; $F(1,46) = 7.4$, MSe = 42.7, $P < 0.01$).

There were more errors for the problems in which units had to be added, and in particular for those that involved a carry-over operation. The latter is a well established finding in the literature (e.g. Timmers and Claeys, 1990). More errors were also made when problems were presented in the verbal than when they were presented in the Arabic format. Finally, the presentation order in which the larger number preceded the smaller number elicited more mistakes. Interestingly for the rest of the discussion, although there was no significant interaction between language group and order of presentation ($F(1,46) < 1$), the higher error rate for T&U relative to U&T problems tended to be more pronounced for the Dutch-speaking participants (9.8 vs. 7.8%) than for the French-speaking participants (9.0 vs. 8.1%).

The reaction times of the correct trials with a good time measurement as a function of the different variables are shown in Table 1 (in general, there were 5% bad time registrations due to premature voice key triggering; these were more likely in conditions with long response times). The $2 \times 3 \times 2 \times 2$ analysis of variance yielded significant main effects for all variables: language group (French = 1536 ms, Dutch = 1411 ms; $F(1,46) = 4.6$, MSe = 488 923, $P < 0.05$), problem type

Table 1
Reaction times and percentages of error (in brackets) as a function of language, order of presentation, problem type, and format; experiment 1, oral response

		French			Dutch		
		T&U	U&T	Difference	T&U	U&T	Difference
T + U	Arabic	929 (3.7)	1039 (3.1)	110	951 (3.1)	1025 (3.1)	74
	Verbal	1235 (6.4)	1353 (5.0)	118	1087 (5.0)	1089 (3.7)	2
TU + Unc	Arabic	1096 (6.0)	1152 (5.0)	56	1202 (8.5)	1151 (4.7)	−51
	Verbal	1704 (8.6)	1732 (8.3)	28	1592 (11.4)	1569 (8.8)	−23
TU + Uc	Arabic	1564 (11.6)	1650 (9.1)	86	1504 (15.0)	1502 (11.6)	−2
	Verbal	2457 (18.0)	2520 (17.9)	63	2108 (15.7)	2154 (14.7)	46

(T + U = 1088 ms, TU + Unc = 1400 ms, TU + Uc = 1932 ms; $F(2,92)$ = 369.5, MSe = 94 670, $P < 0.01$; post hoc Newman–Keuls tests showed that each comparison was significant), format (Arabic = 1230 ms, verbal = 1717 ms; $F(1,46)$ = 191.9, MSe = 177 478, $P < 0.01$), and order of presentation (T&U = 1452 ms, U&T = 1495 ms; $F(1,46)$ = 16.1, MSe = 15 944, $P < .01$). Furthermore, all but one first-order interactions were significant: language group × problem type ($F(2,92)$ = 4.7, MSe = 94 670, $P < 0.02$), language group × format ($F(1,46)$ = 9.7, MSe = 177 478, $P < 0.01$), language group × order of presentation ($F(1,46)$ = 10.7, MSe = 15 944, $P < 0.01$), problem type × format ($F(2,92)$ = 117.3, MSe = 30 972, $P < 0.01$), and problem type × order of presentation ($F(2,92)$ = 4.1, MSe = 16 283, $P < 0.01$).

These results show that the Dutch-speaking participants solved the addition problems faster than the French-speaking participants. This was especially true for problems presented in the verbal format. In line with previous studies, additions were more difficult with problems presented in the verbal format than with problems presented in the Arabic format. The difference between both formats increased for the more difficult problems. Finally, and most importantly for the present discussion, there was an interaction between language group and order of presentation, which to some extent seemed to depend on the type of problem. To get a clearer picture of the latter finding, separate analyses were run for the different problem types. As there were no reliable interaction effects for the percentages of errors, the additional analyses were restricted to the reaction times (RTs).

2.2.1. Ten + unit problems

A $2 \times 2 \times 2$ ANOVA with the variables language group, format, and order of presentation revealed significant F-values for all main and interaction effects. In general, the French-speaking participants were slower than the Dutch-speaking ones (1139 vs. 1038 ms; $F(1,46)$ = 9.16, MSe = 53 536, $P < 0.01$), but this difference was entirely due to the problems presented in the verbal mode (1294 vs. 1088 ms). Both groups performed similarly on the problems presented in the Arabic mode (French = 984 ms; Dutch = 988 ms; interaction language × format: $F(1,46)$ = 61.06, MSe = 4553, $P < 0.01$). The main effect of format was significant ($F(1,46)$ = 79.87, MSe = 25 277, $P < 0.01$) because both groups needed more time to respond to verbal problems than to Arabic problems (as indicated above, the difference due to the format was larger for the French-speaking than for the Dutch-speaking participants).

More importantly, the order in which the operands had been presented had a significant effect ($F(1,46)$ = 61.06, MSe = 4553, $P < 0.01$) that differed as a function of the language group (interaction language × order: $F(1,46)$ = 15.29, MSe = 4553, $P < 0.01$), the format (interaction order × format: $F(1,46)$ = 4.60, MSe = 2692, $P < 0.05$), and the combination of language and format (interaction order × language × format: $F(1,46)$ = 7.28, MSe = 7.28, MSe = 2692, $P < 0.01$). To clarify this pattern of findings, separate ANOVAs with format and order as repeated measures were run on the data of the two language groups.

RTs of the French-speaking subjects varied significantly as a function of format

(Arabic = 984 ms; verbal = 1294 ms; $F(1,23) = 56.1$, MSe = 41 184, $P < 0.01$), and order (T&U = 1082; U&T = 1196; $F(1,23) = 76.3$, MSe = 4099, $P < 0.01$). There was no interaction between both variables ($F < 1$) (see upper left panel of Table 1).

RTs of the Dutch-speaking subjects also varied significantly as a function of format (Arabic = 988 ms; verbal = 1088 ms; $F(1,23) = 25.6$, MSe = 9371, $P < 0.01$) and order (T&U = 1019; U&T = 1057; $F(1,23) = 6.9$, MSe = 5007, $P < 0.05$), but in addition there was a significant interaction between both variables ($F(1,23) = 13.1$, MSe = 2413, $P < 0.01$). This was due to the fact that the effect of order was only present for problems presented in Arabic mode, as can be seen in the upper right panel of Table 1.

In summary, the French-speaking subjects experienced an advantage of 114 ms when the problems were presented in the T&U order (20 + 4) than when they were presented in the U&T order (4 + 20). This effect was the same whether problems were presented in the Arabic or verbal format (110 vs. 118 ms). In contrast, the Dutch-speaking subjects only showed a 74 ms advantage for the T&U order when the problems were presented in Arabic code. For the verbal code, there was no difference between the T&U and U&T order (2 ms). ANOVAs with the variables language and order revealed a significant interaction between both factors for the verbal problems ($F(1,46) = 15.9$, MSe = 1539, $P < 0.01$) and a nearly significant interaction for the Arabic problems ($F(1,46) = 3.7$, MSe = 2106, $P < 0.07$).

2.2.2. Ten-unit + unit without carry-over problems

RTs of correct responses differed significantly as a function of format ($F(1,46) = 180.3$, MSe = 66 308, $P < 0.01$), language × format ($F(1,46) = 6.6$, MSe = 66 308, $P < 0.05$), and language × order ($F(1,46) = 9.6$, MSe = 7813, $P < 0.01$). The interaction between language and format was due to the fact that both language groups had the same response times for Arabic problems (French = 1124 ms, Dutch = 1176 ms; $F(1,46) < 1$), whereas RTs for verbal problems were considerably longer for the French-speaking (1718 ms) than for the Dutch-speaking (1581 ms; $F(1,46 = 7.3$, $P < 0.01$) subjects. The interaction between language and order was caused by the fact that T&U problems (24 + 1) were solved faster than U&T (1 + 24) problems for the French-speaking (1400 and 1442 ms, respectively; $F(1,46) = 4.2$, $P < 0.05$), whereas the reverse was true for the Dutch-speaking (1397 and 1360 ms, respectively; $F(1,46) = 5.4$, $P < 0.05$; see the middle part of Table 1). Separate ANOVAs on the two language groups showed non-significant interactions between the order effect and format ($F < 1.2$).

2.2.3. Ten-unit + unit with carry-over problems

The analysis of variance on the RTs (see lower part of Table 1) resulted in significant main effects of language ($F(1,46) = 5.3$, MSe = 48 6245, $P < 0.05$) and format ($F(1,46) = 185.0$, MSe = 14 7838, $P < 0.01$), and an interaction between format and language ($F(1,46) = 5.3$, MSe = 14 7838, $P < 0.05$). The interaction was due to the fact that, just as for the other types of problems, the French-speaking subjects were slower than the Dutch-speaking subjects for the problems

presented in the verbal format (2489 vs. 2131 ms; $F(1,46) = 6.2, P < 0.05$) whereas there was no great difference between both groups for the problems presented in the Arabic format (French = 1607 ms; Dutch = 1503 ms; $F(1,46) = 1.8, P > 0.15$). Separate ANOVAs for both language groups did not yield other significant effects apart from those covered by the general ANOVA.

2.3. Discussion

Experiment 1 was designed to find out whether we could find differences in arithmetical performance due to the native language of the person. More precisely, we wanted to find out whether the reversed number naming of two-digit numbers in Dutch implied better performance for problems of the type 4 + 20 (four-and-twenty) than for problems of the type 20 + 4, and whether performance on these problems would differ from that of French-speaking participants. Finding such a Whorfian effect would indicate that mathematical operations are not completely impervious to language influences, as defended in some models of mathematical cognition.

As can be seen in the upper part of Table 1, there was indeed a language difference for the T + U problems. However, the difference depended on the input format. In particular, when problems were presented in Arabic code, the French-speaking subjects showed a 110 ms advantage for the T&U order (20 + 4) compared with the U&T order (4 + 20). The Dutch-speaking subjects showed a similar but smaller 74 ms advantage for the T&U order. The T&U advantage for the French-speaking subjects was the same (118 ms) when problems were presented in the verbal code, as could be expected on the basis of the similar structure of two-digit numbers in Arabic and verbal modality. In contrast, the T&U advantage for verbally presented problems was completely absent (2 ms) for the Dutch-speaking subjects. It is noteworthy, however, that we did not obtain the reversed pattern here (i.e. faster U&T solution times), as might have been expected on the basis of the Whorfian hypothesis. Implications of these findings will be addressed in Section 4.

In addition to the language effect we were testing, two other language differences emerged from our study. First, the French-speaking subjects were faster to add the TU + U without carry-over problems (24 + 1) than the U + TU without carry-over problems (1 + 24), whereas the reverse was true for the Dutch-speakers. This effect did not depend on the input modality (i.e. it was the same for the Arabic and the verbal code; see the middle part of Table 1). Second, the processing cost of the verbal presentation modality was larger for the French-speaking than for the Dutch-speaking subjects.

One interpretation of the language difference for the TU + Unc problems is that it is a genuine Whorfian effect and caused by the facts that (i) addition processes are based on verbal codes, and (ii) subjects experience an advantage when the digits of the operands succeed one another. According to this interpretation, it is easier to solve 'twenty-four plus two' and 'two plus four-and-twenty' because the units are next to one another, in contrast to the cases 'two plus twenty-four' and 'four-and-twenty plus two' where the ten interferes.

Another interpretation, however, may be that subjects alter their addition strategies depending on the output format. More specifically, the Dutch-speakers try to get access to the unit of the response first, because they can start programming the pronunciation of the answer as soon as the value of the unit is known. In contrast, the French-speakers have to capitalise on the value of the ten, which they must know before response execution can be started. This interpretation would also explain why the order effect largely disappears for the problems with a carry-over operation (where the unit of the response is not simply the sum of the units of the operands), and why the cost of the verbal presentation format was larger for the French-speaking than for the Dutch-speaking subjects (because the response times of the Dutch-speaking subjects did not reflect the calculation of the whole solution). If the second interpretation is correct, a straightforward prediction follows: the U + TU without carry-over advantage will turn into a TU + U advantage for the Dutch-speaking subjects if the subjects have to output the ten of the response before the unit. This idea is tested in the next experiment.

3. Experiment 2

Experiment 1 returned a remarkable effect because the largest difference between French and Dutch was not found for the T + U (20 + 4) vs. U + T (4 + 20) problems, as expected, but for the TU + U (24 + 1) vs. U + TU (1 + 24) problems. The present experiment tried to find out whether this language difference is due to processes involved in the addition operation, or rather to output requirements. This was done by asking French and Dutch-speaking subjects not to pronounce the correct response, but to type in the Arabic representation on the keyboard. In this way, for both language groups the value of the ten had to be known before the solution could be entered.

3.1. Method

3.1.1. Subjects
Twenty-four new native French-speaking students from the Université Catholique de Louvain and 24 new Dutch-speaking students from the Katholieke Universiteit Leuven participated on a voluntary basis. As in the previous experiment, they were ignorant about the research hypotheses.

3.1.2. Procedure
The procedure was the same as in experiment 1, except for the fact that subjects had to press on a button to see the stimulus and had to type the response on the keyboard instead of pronouncing it. At the beginning of a trial, subjects had to press on an external button connected to the game port in order to see the stimulus. The stimulus remained on the screen as long as the subject pressed on the button and disappeared as soon as he/she released the button. In this way, we had two dependent variables, namely the time when the button was released and the time when the first

key of the keyboard was pressed. We opted for this procedure because it allowed us to get rid of differences in response time due to the search for the correct digit key on the keyboard.

3.2. Results

An overall $2 \times 3 \times 2 \times 2$ analysis of variance involving the variables language, problem type, format, and order of presentation, was run on the percentage of errors (see Table 2). This yielded a main effect of language group (French = 3.5%, Dutch = 5.9%; $F(1,46) = 12.9$, MSe = 63.5, $P < 0.01$), problem type (T + U = 2.1%, TU + Unc = 3.7%, TU + Uc = 8.2%; $F(2,92) = 52.3$, MSe = 36.3, $P < 0.01$; post hoc Newman–Keuls tests indicated that all comparisons were significant), and format (Arabic = 3.9%, verbal = 5.4%; $F(1,46) = 9.8$, MSe = 34.3, $P < 0.01$). There was also a significant interaction between format type and order of presentation (Arabic: T&U = 3.4%, U&T = 4.4%; verbal: T&U = 5.8%, U&T = 5.0%; $F(1,46) = 5.4$, MSe = 21.6, $P < 0.01$). Planned contrasts indicated that the order effect was significant for the Arabic problems ($F(1,46) = 5.1$, $P < 0.05$), but not for the verbal problems ($F(1,46) = 1.6$, $P > 0.20$). Although the second-order interaction was not significant, the interaction effect between format type and order of presentation was larger for the Dutch speaking than for the French speaking subjects (see Table 2 for the data).

As in experiment 1, the French-speaking subjects produced less errors than the Dutch-speaking subjects. Error rates also differed as a function of the presentation format of the addends, with a higher percentage of errors for the verbal format than for the Arabic format. Percentages of errors increased with the difficulty of the problem: The lowest error rate was observed for the T + U problems and the highest error rate for the TU + Uc problems. Contrary to experiment 1 where a main effect of order of presentation was observed, here, the order effect depended on the presentation format: there was a significant advantage of the T&U order for the Arabic problems, and a tendency towards the opposite pattern for the verbal problems.

The same $2 \times 3 \times 2 \times 2$ analysis applied to the button release times (Table 2) returned significant main effects for all four variables: language group (French = 1701 ms, Dutch = 1447 ms; $F(1,46) = 4.4$, MSe = 2 172 165, $P < 0.01$), problem

Table 2
Presentation button release times and percentages of error (in brackets) as a function of language, order of presentation, problem type, and format; experiment 2, typing response

		French			Dutch		
		T&U	U&T	Difference	T&U	U&T	Difference
T + U	Arabic	925 (1.5)	982 (1.5)	57	807 (2.2)	870 (1.9)	63
	Verbal	1297 (1.7)	1378 (5.0)	81	1136 (3.8)	1135 (2.2)	−1
TU + Unc	Arabic	1106 (0.7)	1154 (2.2)	48	966 (2.2)	1065 (3.8)	99
	Verbal	1888 (3.6)	1952 (2.3)	64	1745 (5.2)	1805 (6.7)	60
TU + Uc	Arabic	1899 (4.3)	1877 (5.6)	−22	1480 (4.9)	1462 (7.7)	−18
	Verbal	2940 (7.8)	3010 (8.2)	70	2401 (10.0)	2450 (11.3)	49

type (T + U = 1066 ms, TU + Unc = 1460 ms, TU + Uc = 2190 ms; $F(2,92)$ = 136.4, MSe = 457 649, $P < 0.01$; all comparisons were significant according to the Newman–Keuls test), format (Arabic = 1216 ms, verbal = 1928 ms; $F(1,46)$ = 237.6, MSe = 307 106, $P < 0.01$), order of presentation (T&U = 1549 ms, U&T = 1595 ms; $F(1,46)$ = 7.9, MSe = 38 390, $P < 0.01$). In addition, there was an interaction between language group and problem type ($F(2,92)$ = 4.0, MSe = 457 649, $P < 0.05$), due to the fact that the difference between the Dutch and the French-speaking subjects increased for the more difficult problems, and an interaction between format and problem type ($F(2,92)$ = 99.1, MSe = 57 469, $P < 0.01$), due to the fact that the difference between the Arabic and the verbal mode increased for the more difficult problems.

Overall, the Dutch-speaking subjects released the presentation button faster than the French-speaking subjects but made more errors. As in experiment 1, there were huge effects of problem type and format, which interacted. More importantly, we again found a very reliable effect of the order of presentation: problems that had the largest number presented first (e.g. 21 + 4) could be solved faster than problems with the smallest number first (e.g. 4 + 21). Contrary to experiment 1, this order effect was not involved in any significant interaction effect. However, to get a better idea of what was going on, we calculated separate analyses for the different problem types. Because no additional information was present in the error data, the analyses were restricted to the reaction times.

3.2.1. Ten + unit problems

Button release time varied significantly as a function of language group ($F(1,46)$ = 7.9, MSe = 153 100, $P < 0.01$), format ($F(1,46)$ = 147.1, MSe = 37 805, $P < 0.01$) and order of presentation ($F(1,46)$ = 19.8, MSe = 6094, $P < 0.01$). Furthermore, there was a significant interaction between the three variables ($F(1,46)$ = 5.0, MSe = 4884, $P < 0.05$). A look at Table 2 shows that this was due to the deviating order effect for the Dutch-speaking subjects when the problems were presented in the verbal mode. Whereas the order effect was the same for the French and the Dutch-speaking subjects in the Arabic mode (57 and 63 ms, respectively; $F(1,46) < 1$), it differed strongly in the verbal modality (81 vs. −1 ms; $F(1,46)$ = 4.8, $P < 0.05$).

3.2.2. Ten-unit + unit without carry-over problems

The button release time varied significantly as a function of format ($F(1,46)$ = 280.4, MSe = 102 800, $P < 0.01$) and order of presentation ($F(1,46)$ = 12.1, MSe = 18 089, $P < 0.01$). There was no significant effect of language group ($F(1,46)$ = 1.7) and all other F-values were smaller than 1.

3.2.3. Ten-unit + unit with carry-over problems

Button release time differed as a function of language group ($F(1,46)$ = 4.6, MSe = 2 455 796, $P < 0.05$) and format ($F(1,46)$ = 177.6, MSe = 281 445, $P < 0.01$). All other F-values were smaller than 1.1.

Finally, to ensure that none of the important effects were obscured by differences

in the time period between the button release that made the problem disappear and the first key press on the computer keyboard (see Section 3.1), we also ran a $2 \times 3 \times 2 \times 2$ analysis of variance on the time lapse between the button release and the first key press. This time differed significantly as a function of problem type (T + U = 599 ms, TU + Unc = 677 ms, TU + Uc = 758 ms; $F(2,92) = 28.6$, MSe = 42 261, $P < 0.01$; all comparisons were significant according to the New-man–Keuls test), mimicking the differences in the button release time. There was also a significant interaction between language group and format ($F(1,46) = 4.9$, MSe = 53 429, $P < 0.05$), because for the French-speaking subjects the time lapse tended to be shorter for problems presented in the verbal than in the Arabic modality (686 vs. 723 ms; $F(1,46) = 1.8$, $P < 0.20$), whereas the Dutch-speaking subjects showed the opposite pattern of results (675 vs. 627 ms, respectively; $F(1,46) = 3.1$, $P < 0.10$). No other effect was significant, and a look at the values in the different experimental cells indicated that no reliable finding reported in Table 2 was offset by a spill-over effect on the typing time. This was further verified by looking at the keypress times (i.e. the latency between the onset of the stimulus and the first key-press on the numerical path of the keyboard): all effects due to the order of presentation remained the same.

3.3. Discussion

There were three important findings in experiment 2. First, the unexpected language difference for the TU + Unc problems in experiment 1 disappeared when subjects had to type the solution on a keyboard instead of pronouncing it. Both the Dutch and the French-speaking subjects now showed a reliable T&U advantage for this type of problem. This indicates that the language difference found for the TU + U problems in experiment 1 was not due to a language influence at the mathematical addition stage, but was caused by differences in response strategies used by the subjects as a function of output requirements. In particular, when the answer had to begin with the unit value, subjects seemed to be able to start out-putting this value before the complete solution was calculated. In this respect, it is interesting to note that the time differences between the Arabic and the verbal presentation format in experiment 2 were very similar for both language groups and were more in line with those of the French-speaking subjects of experiment 1 than with those of the Dutch-speaking subjects (i.e. considerably larger format differences for the more difficult types of problems; see Table 2).

Second, the new data show that the order effect of the T + U problems presented in the Arabic modality is the same for the Dutch-speaking subjects (a 63 ms advantage for the T&U order over the U&T order) as for the French-speaking subjects (an advantage of 57 ms). In experiment 1, the nearly significant difference between the 110 ms advantage of the French-speaking subjects and the 74 ms advantage of the Dutch-speaking subjects could have been interpreted as evidence for the Whorfian hypothesis. However, this interpretation can no longer be sustained given the lack of a language effect in the present experiment where both the presentation format (i.e. the Arabic format) and the response order (i.e. starting with the value of the ten)

were matched for the two language groups. More in general, it may be remarked that the order effects of the two language groups for all types of problems, except one, resembled each other much more in experiment 2 than in experiment 1 (i.e. compare Tables 1 and 2).

Finally, the overall more similar pattern of the data for the French and Dutch-speaking subjects in experiment 2 did not attenuate the finding of experiment 1 that the order effect completely disappeared for the Dutch language group when T + U problems are presented in verbal format. Whereas the French-speaking subjects showed a reliable 81 ms advantage for the T&U order, no comparable difference was found for the Dutch-speaking subjects (−1 ms). Implications of these findings for the Whorfian hypothesis and for the theories of numerical cognition will be outlined below.

4. General discussion

Two recent theoretical moves have refuelled the interest in the Whorfian hypothesis among experimental psychologists. The first is that a Whorfian effect should not be considered as a manifestation of linguistic determinism, but rather as an indication of the extent to which a particular part of the cognitive system can be penetrated by language processes or is a language-independent, encapsulated module. The second insight is that a Whorfian effect does not imply that a particular function cannot be performed in a particular language, but rather that some functions may be favoured by some languages and other functions by other languages (Hunt and Agnoli, 1991). Instead of looking for all-or-none effects, researchers should search for differences in processing costs related to the language of the subject. In experimental psychology, most of the time this is translated into significant, language-dependent differences in response times. A further restriction to this approach is that researchers should verify that the language difference they obtain is related to the mental representation of the stimulus and not to peripheral input or output requirements.

In this article we have investigated whether we can find a Whorfian effect in an over-learned mathematical operation (the addition of two Arabic or verbal operands). Numerical cognition is an interesting research topic to look for Whorfian effects because quite a lot of theorising within this field deals with the questions (i) whether different numerical notations (Arabic, verbal, analogue) are converted to the same input-independent code before further processing takes place, and (ii) to what extent this code (or codes) hinges on language processes. So, a real language effect would be very informative about the correctness of the different mathematical models proposed so far (see Section 1).

To investigate the possibility of a Whorfian effect in numerical cognition, we made use of a characteristic in the Dutch number naming system. Given that the Dutch verbal representation of a two-digit number reverses the position of the unit and the ten, we wondered whether this would have an effect on the solution times for addition problems that involve the constituents of a TU number. A similar manip-

ulation has recently been used by Noël and Seron (1997) (see Section 1.5), but with less well-known stimulus materials (Roman numerals and large numbers).

At first sight, experiment 1 offered compelling evidence for language influences on mental addition operations. Comparing the performance of French- and Dutch-speaking subjects, we were able to show a series of language dependent effects on the solution times of additions involving a single-digit numeral and a two-digit numeral (e.g. 20 + 4, or 4 + 20). To summarise, in a language that saves the order of the tens and the units in the number names (French), there was a pronounced advantage when the two-digit number preceded the single-digit number. This was true for the Arabic as well as the verbal input, and for problems involving the addition of a ten + unit (20 + 4) or – to a lesser extent – a ten-unit + unit without carry-over (24 + 1). In contrast, when the language reversed the names of the tens and the units (Dutch), the advantage of having the largest numeral first decreased significantly with ten + unit Arabic problems, disappeared totally for similar ten + unit word problems, and even reversed for the ten-unit + unit without carry-over problems in both presentation formats. In general, for all types of problems, the advantage of having the two-digit number first was smaller for Dutch-speaking participants than for French-speaking participants (see Table 1).

However, in a subsequent experiment in which we asked subjects to type in the answer, so that the Dutch-speaking subjects had to respond in the same way as the French-speaking subjects (i.e. by starting the response with the tens rather than with the units), all differences but one disappeared (see Table 2). First, the results with the ten-unit + unit problems were now in line with those of the French-speaking subjects. That is, for additions with a ten-unit and no carry-over operation, there was an advantage of the ten-unit + unit order (21 + 4) over the unit + ten-unit order (4 + 21), suggesting that this language difference in experiment 1 had been caused by output requirements rather than by language specific addition processes. What probably happened, is that in the naming task the Dutch-speaking subjects were able to start outputting the sum of the units before the complete solution was calculated (e.g. four + one-and-twenty = FIVE-(and twenty)). This interpretation also explains (i) why the language difference was not so strong for the problems requiring a carry-over operation (in which case it was not easy to discard the tens), and (ii) why the penalty for verbally presented numerals was smaller for the Dutch-speaking than for the French-speaking subjects when the solution had to be pronounced (but not when it had to be typed in).

The most important and interesting data concerned the ten + unit problems. For these problems, a language difference was found for the verbal presentation modality both in experiments 1 and 2. However, in the Arabic modality, the language effect was considerably weaker in experiment 1 (naming) and completely absent in experiment 2 (typing). The latter finding is of particular importance because the condition ten + unit/Arabic presentation/typed response provided the best test for a genuine Whorfian effect in numerical cognition. It was the only condition in which both the presentation format and the output requirements were the same for the two language groups (so that no peripheral input or output processes could account for an observed language difference). Failing to find evidence for an effect in this condition

due to the subject's native language, therefore, means that arithmetical fact retrieval for the addition problems '4 + 20' and '20 + 4' does not differ between a person who customarily says 'four-and-twenty' and a person who is used to saying 'twenty-four'.

Such a finding is not predicted by models that assume mathematical operations to be based on internal verbal codes, like Dehaene's triple code model (Dehaene, 1992). In addition, according to this model, the effect of language should be the same in the verbal and the Arabic condition, as both 'twenty + four' and '20 + 4' are transcoded into the corresponding auditory–verbal representation before the sum is retrieved. A possible way out for the model would be to assume that the differences between the verbal and the Arabic condition are entirely due to peripheral strategies[2]. According to this view, subjects would use the following strategy: when the two-word stimulus they see (e.g. 'twenty + four') forms a syntactically correct two-digit numeral ('twenty-four'), they short-cut the addition process and simply name what they see (at least on some proportion of the trials). This predicts that French-speaking subjects should be relatively faster with 'vingt + quatre', and Dutch-speaking subjects relatively faster with 'vier + twintig'. As the same strategy does not apply to '20 + 4' (204?) nor to '4 + 20' (420?), this would explain the absence of an effect with Arabic notation. A problem with this explanation, however, is that it is not clear how it can be extended to the typing experiment, where we found a similar pattern of results, despite the fact that subjects had to type in exactly the same response.

In contrast, our findings are well in line with predictions of the encoding complex view proposed by Campbell (1994). According to this model, mathematical operations are performed on specific codes that depend on the input format. Thus, a problem presented in the Arabic format is likely to activate the same non-verbal codes for both French- and Dutch-speaking subjects, and no language difference is expected. However, a problem presented in the verbal format is assumed to activate predominately language-related representations that could very well include language particular properties of the number names. Hence, a Whorfian effect is predicted here, as we indeed observed.

Although our results could easily be explained with Campbell's encoding complex model (Campbell, 1994), we hesitate to do so, because the model predicts Whorfian effects for other mathematical operations that are not influenced by language differences (Noël et al., 1997). For instance, Campbell (1994) used the encoding complex idea to explain why in multiplication operations operand intrusion errors are more frequently produced in response to problems presented in word format than to problems presented in digit format. Operand intrusion errors are errors in which one of the problem's operands intrudes in the response (e.g. $4 \times 6 = 42$, 36,...). According to Campbell (1994), these errors are an indication of interactions between arithmetical fact retrieval and reading processes (which are more likely for verbally presented problems than for problems presented in the Arabic format). Faced with a multiplication problem, people would not only activate

---

[2]The authors thank Stanislas Dehaene for pointing them to this possibility.

the corresponding product but also, unintentionally, reading-based associations which have the left operand encoded as the decade number and the right operand encoded as the unit number (e.g. 4×6 activates 'forty-six'). This explains why in most of the operand intrusion errors the intruder keeps its position from the original formulation (e.g. the error $4 \times 6 = 42$ is more likely than the error $4 \times 6 = 54$, because in the former case the number 4 has the same position in the answer as in the problem). However, Noël et al. (1997) showed that this interpretation did not work. They reasoned that if Campbell's explanation was correct, a difference in error pattern should be found between French- and Dutch-speaking individuals. If a Dutch-speaking person reads 48 as 'eight-and-forty', then he should produce more errors like $4 \times 6 = 54$ (four times six equals four-and-fifty) than errors like $6 \times 4 = 54$ (six times four equals four-and-fifty). However, this did not happen: The operand intrusions were exactly the same in French and in Dutch. On the basis of this result, Noël et al. argued that the interactions between reading processes and multiplication solutions did not occur at the stage of arithmetical fact retrieval but at a later, output, stage (cf. their cascade model).

A similar interpretation could be given for the present results. Indeed, all too often, it has been assumed that a numerical model based on abstract codes (McCloskey, 1992), implies that the surface information of the input is lost as soon as the input is converted to the abstract code. This would indeed be the case in an exaggerated, purely serial model where the information strictly goes from one stage to the other. However, there is evidence that numerical input is simultaneously encoded in more than one way and processed in parallel, so that not only the information from the arithmetical fact retrieval system reaches the response output buffer but also parts of the input stimulus and probably other types of information (Noël et al., 1997). The need for such additional information can easily be inferred from tasks where bilinguals have to name the solution of a mathematical problem in one of their languages. According to the exaggerated interpretation of McCloskey's model, bilinguals would no longer know which output language to use once the numerals have been converted to the abstract identities.

The fact that the surface information of the input stimulus is not completely lost, was also demonstrated by Noël and Seron (1997), who put forward the notion of intermediate representations (see Section 1) and published results that were very similar to ours. In particular, they claimed that the effect of the intermediate representations is strongest when there is a perfect match between the operation to be performed and the information disclosed by the input stimulus. So, they observed faster verification times for $VII = 5 + 2$ but not for $VII = 6 + 1$, just as we observed a language effect for the verbal $T + U$ problems, but not for the verbal $TU + U$ problems. Given these observations, our finding that 'four + twenty' did not take longer to solve than 'twenty + four' for a Dutch-speaking participant, is more likely to be due to the fact that the former syntax interferes less with the response 'four-and-twenty' than to the fact that a different arithmetical fact retrieval system is accessed for problems presented in the Arabic and in the verbal mode.

All in all, instead of showing a Whorfian effect, our study has demonstrated how careful one must be in interpreting a language difference in a numerical task as the

result of a difference in the semantic number system. Only by carefully controlling all input and output factors, is it possible to disentangle a real Whorfian effect from peripheral language differences. As such, our results add evidence to the idea that the numerical system is largely autonomous of the language system (except maybe during the acquisition phase). As indicated in Section 1, this of course does not mean that Whorfian effects cannot exist in other areas of human cognition.

Acknowledgements

References

Ashcraft, M.H., 1992. Cognitive arithmetic: a review of data and theory. Cognition 44, 75–106.

Au, R., 1983. Chinese and English counterfactuals: the Sapir–Whorf hypothesis revisited. Cognition 15, 155–187.

Baddeley, A.D., Thomson, N.K., Buchanan, M., 1975. Word-length and the structure of short-term memory. Journal of Verbal Learning and Verbal Behavior 14, 575–589.

Berlin, B., Kay, P., 1969. Basic Color Terms: Their Universality and Evolution. University of California Press, Berkeley, CA.

Blankenberger, S., Vorberg, D., 1997. The single-format assumption in arithmetic fact retrieval. Journal of Experimental Psychology: Learning, Memory, and Cognition 23, 721–738.

Bloom, A.H., 1981. The Linguistic Shaping of Thought: A Study in the Impact of Language on Thinking in China and the West. Erlbaum, Hillsdale, NJ.

Bovens, N., Brysbaert, M., 1990. IBM PC/XT/AT and PS/2 Turbo Pascal timing with extended resolution. Behavior Research Methods, Instruments, and Computers 22, 332–334.

Brown, R., 1976. Reference: in memorial tribute to Eric Lenneberg. Cognition 4, 125–153.

Brysbaert, M., 1995. Arabic number reading: on the nature of the numerical scale and the origin of phonological recoding. Journal of Experimental Psychology: General 124, 434–452.

Campbell, J.I.D., 1994. Architectures for numerical cognition. Cognition 53, 1–44.

Campbell, J.I.D., Clark, J.M., 1988. An encoding complex view of cognitive number processing: comment on McCloskey, Sokol, and Goodman (1986). Journal of Experimental Psychology: General 117, 204–214.

Dehaene, S., 1992. Varieties of numerical abilities. Cognition 44, 1–42.

Deloche, G., Seron, X., 1982. From one to 1: an analysis of a transcoding process by means of neuropsychological data. Cognition 12, 119–149.

Ellis, N.C., Hennely, R.A., 1980. A bilingual word-length effect: implications for intelligence testing and the relative ease of mental calculation in Welsh and English. British Journal of Psychology 7, 43–51.

Gallistel, C.R., Gelman, R., 1992. Preverbal and verbal counting and computation. Cognition 44, 43–74.

Geary, D.C., 1996. International differences in mathematical achievement: their nature, causes, and consequences. Current Directions in Psychological Science 5, 133–137.

Gerrig, R.J., Banaji, M.R., 1994. Language and thought. In: Sternberg, R. (Ed.), Handbook of Perception and Cognition: Thinking, Vol. 12. Academic Press, Orlando, FL.

Greenberg, J.H., 1978. Generalizations about numeral systems. In: Greenberg, J.H. (Ed.), Universals of Human Language: Vol. 3, Word Structure. Stanford University Press, Stanford, CA, pp. 249–295.

Hardin, C., Banaji, M.R., 1993. The influence of language on thought. Social Cognition 11, 277–308.

Heider, E.R., 1972. Universals in color naming and memory. Journal of Experimental psychology 93, 10–20.

Heider, E.R., Olivier, D.C., 1972. The structure of the color space in naming and memory for two languages. Cognitive Psychology 3, 337–354.

Hill, J.H., Mannheim, B., 1992. Language and world view. Annual Review of Anthropology 21, 381–406.

Hoffman, C., Lau, I., Johnson, D.R., 1986. The linguistic relativity of person cognition. Journal of Personality and Social Psychology 51, 1097–1105.

Hoosain, R., 1986. Language, orthography and cognitive processes: Chinese perspectives for the Sapir–Whorf hypothesis. International Journal of Behavioral Development 9, 507–525.

Hunt, E., Agnoli, F., 1991. The Whorfian hypothesis: a cognitive psychology perspective. Psychological Review 98, 377–389.

Hurford, J.R., 1987. Language and Number. Basil Blackwell, Oxford.

Kay, P., Kempton, W., 1984. What is the Sapir–Whorf hypothesis?. American Anthropologist 86, 65–79.

Kolers, P.A., 1968. Bilingualism and information processing. Scientific American 218, 78–86.

Kolinsky, R., Morais, J., Brito Mendes, C., 1990. Embeddedness effects on part verification in children and unschooled adults. Psychologica Belgica 30, 49–64.

Liu, L.G., 1985. Reasoning counterfactually in Chinese: are there any obstacles?. Cognition 21, 239–270.

Lucy, J.A., Shweder, R.A., 1979. Whorf and his critics: linguistic and non-linguistic influences on color memory. American Anthropologist 81, 581–607.

Marsh, L.G., Maki, R.H., 1976. Efficiency of arithmetical operations in bilinguals as a function of language. Memory and Cognition 4, 216–230.

McClain, L., Shih Huang, J.Y., 1982. Speed of simple arithmetic in bilinguals. Memory and Cognition 10, 591–596.

McCloskey, M., 1992. Cognitive mechanisms in numerical processing: evidence from acquired dyscalculia. Cognition 44, 107–157.

McCloskey, M., Macaruso, P., 1995. Representing and using numerical information. American Psychologist 50, 351–363.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., Amiel-Tison, C., 1988. A precursor of language acquisition in young infants. Cognition 29, 143–178.

Miller, K.F., 1996. Origins of quantitative competence. In: Gelman, R., Au, T.K.F. (Eds.), Perceptual and Cognitive Development. Academic Press, San Diego, CA, pp. 213–243.

Miller, K.F., Zhu, J., 1991. The trouble with teens: accessing the structure of number names. Journal of Memory and Language 30, 48–68.

Miura, I.T., Okamoto, Y., Kim, C.C., Steere, M., Fayol, M., 1993. First graders' cognitive representation of number and understanding of place value: cross-national comparisons – France, Japan, Korea, Sweden, and the United States. Journal of Educational Psychology 85, 24–30.

Morais, J., Bertelson, P., Cary, L., Alegria, J., 1986. Literacy training and speech segmentation. Cognition 24, 45–64.

Morais, J., Kolinsky, R., 1994. Perception and awareness in phonological processing: the case of the phoneme. Cognition 50, 287–297.

Naveh-Benjamin, M., Ayres, T.J., 1986. Digit span, reading rate, and linguistic relativity. The Quarterly Journal of Experimental Psychology 38A, 739–751.

Noël, M.P., Fias, W., 1998. Bilingualism and numerical cognition. Psychologica Belgica, in press.

Noël, M.P., Fias, W., Brysbaert, M., 1997. About the influence of the presentation format on arithmetical-fact retrieval processes. Cognition 63, 335–374.

Noël, M.P., Seron, X., 1997. On the existence of intermediate representations. Journal of Experimental Psychology: Learning, Memory, and Cognition 23, 697–720.

Pylyshyn, Z.W., 1984. Computation and Cognition: Toward a Foundation for Cognitive Science. MIT Press, Cambridge, MA.

Read, C., Zhang, Y.F., Nie, H.Y., Ding, B.Q., 1986. The ability to manipulate speech sounds depends on knowing alphabetic writing. Cognition 24, 31–44.

Sapir, E., 1949. In: Mandelbaum, D. (Ed.), Selected Writings of Edward Sapir. University of California Press, Berkeley, CA..

Saunders, B.A.C., van Brakel, J., 1997. Are there non-trivial constraints on colour categorization? Behavioral and Brain Sciences, 20, 167–228.

Saxe, G.B., Posner, J.K., 1983. The development of numerical cognition: cross-cultural perspectives. In: Ginsburg, H.P. (Ed.), The Development of Mathematical Thinking. Academic Press, San Diego, CA, pp. 291–317.

Seron, X., Fayol, M., 1994. Number transcoding in children: a functional analysis. British Journal of Developmental Psychology 12, 281–300.

Shanon, B., 1984. The polyglot mismatch and the monolingual tie: observations regarding the meaning of numbers and the meaning of words. New Ideas in Psychology 2, 75–79.

Timmers, L., Claeys, W., 1990. The generality of mental addition models: simple and complex addition in a decision-production task. Memory and Cognition 18, 310–320.

Towse, J.N., Saxton, M., 1997. Linguistic influences on children's number concepts: methodological and theoretical considerations. Journal of Experimental Child Psychology 66, 362–375.

Werker, J.F., Tees, R.C., 1984. Cross-language speech perception: evidence for perceptual reorganization during the first year of life. Infant Behavior and Development 7, 49–63.

Whorf, B.L., 1956. In: Carroll, J.B. (Ed.), Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf. MIT Press, Cambridge, MA.