Running head: Event Categorisation with Minimal to Maximal L2 Involvement

**Changing Event Categorisation in Second Language Users**

**Through Perceptual Learning**

**Author**

Name: Norbert Vanek

Affiliation: University of York

E-mail: [norbert.vanek@york.ac.uk](mailto:norbert.vanek@york.ac.uk)

Telephone: 0044 1904 323951

Address: Centre for Research in Language Learning and Use (Room D/M/203), Department of Education, University of York, York, YO10 5DD, United Kingdom

ORCID: <https://orcid.org/0000-0002-7805-184X>

This study examines the impact of a second language on event phase categorisation. The aim is to test how strong a boost the L2 system provides when learners are trained to classify events in a new way. The linguistic contrast in focus is grammatical expression of change-of-state events in progress, available in English but far less so in Chinese. Advanced Chinese learners of English received categorisation training in four conditions: action-biased, completion-biased, verbally-distracted and with overt L2 encoding. Experiment 1 tested the extent to which event categorisation is facilitated by the learners’ native language. Experiment 2 compared learning success in L2-based categorisation with and without verbal interference. Experiment 3 measured the benefits of explicit verbalisation for L2-based category learning. The results show a graded impact of L2 grammar on learning success, ranging from the highest with overt verbalisation, followed by categorisation in silence, and the lowest, but not significantly reduced, in learning with verbal interference.

**Keywords**

event categorisation, perceptual learning, linguistic relativity, temporal cognition, grammatical aspect

**Author Note**

Research reported in this study was funded by the ECF-2014-459 grant awarded to Norbert Vanek by the Leverhulme Trust. Many thanks to the anonymous reviewers and the Journal Editor, Pavel Trofimovich, for their helpful feedback on the earlier versions of this article. Appreciation is also due to Alice Cruickshank for her assistance during experiment setup, to Márton Sóskuthy for statistical advice, and to Haoruo Zhang for her help with data collection. Parts of this study were presented at the 2016 UK Cognitive Linguistics Conference.

**Introduction**

Human communication very often involves talking about events, many of which consist of different phases (e.g., *the goalkeeper is kicking a ball away* can be divided into two phases: the ball being kicked and the ball flying away). Despite the ubiquitous nature of such events, languages considerably differ in how events and their phases are encoded (e.g., with focus on the action vs. on the result). Recent research shows that such differences influence event similarity judgments in language-specific ways (Bylund, Athanasopoulos, & Oostendorp, 2013). One possible theoretical explanation is that language and perception are strongly interlinked because linguistic labels activate the visual properties they refer to (Lupyan & Ward, 2013) and thus facilitate perceptual judgments in category recognition (Goldstone & Hendrickson, 2009). This leads to the question of learnability, i.e., whether a new linguistic system can influence perceptual processes of second language (L2) users in predictable ways. With more than one language at their disposal, new evidence suggests that the L2 user’s mind seems to be able to flexibly refocus on specific event features depending on the language of operation (Athanasopoulos et al., 2015). However, the magnitude of impact a new language can have on categorisation is still largely unexplored. Reports from empirical studies to date substantially vary from (1) no impact of first language(L1)-based versus L2-based labels on behavioural performance (even when the task requires overt verbalisation, e.g., Filipović, 2011); to (2), a selective impact only in contexts where covert L2-based labelling is available (e.g., Athanasopoulos & Albright 2016); to (3), a pervasive impact of L2 also in contexts without overt or covert language use (e.g., Dolscheid et al., 2013). This lack of consistent findings can be attributed to a shortage of designs that would enable direct comparisons across tasks in which L2 involvement ranges from minimal to maximal.

The present study addresses this gap through a systematic manipulation of linguistic activity in a perceptual learning context (Kersten et al., 2010). *Perceptual learning* is defined as “changes to an organism’s perceptual system that improve its ability to respond to its environment […] by increasing the attention paid to perceptual dimensions and features that are important” (Goldstone 1998: 585-588).The main aim here is to measure the degree to which behavioral responses of learners change when their L2 highlights different perceptual features than their source language does. Chinese learners of English were recruited for this study. The key contrast between the two languages here is that, for the expression of unfolding change-of-state events, Chinese tends to be more ‘result-oriented’ (a) *The train has arrived at platform one*; while English is more ‘action-oriented’ (b) *The train is arriving at platform one*. This is attributable to grammatical differences (constructions such as (b) are highly infelicitous in Chinese (Xiao & McEnery, 2004)). If one assumes that grammatical differences lead speakers of various languages to different ways of forming event categories, then what exactly is the role of the target language in retraining the mind to form a new event category? Information about changes in categorisation while L2 users are trained to perceive change-of-state events as more action-oriented with or without the involvement of language is still absent. A perceptual learning approach presents a fruitful way to show how strong a boost a new language provides to a new kind of category formation and learning. Looking at learners is beneficial in two ways. It can address language learning questions (e.g., *does increased involvement of L2 labels speed up learning gains?*) as well as inform theories about the interaction between language(s) and thought (e.g., the *attentional learning theory* (Goldstone & Hendrickson, 2009) by examining whether L2 regularities facilitate discrimination between L2-based categories).

**Linguistic relativity effects in categorisation**

Few topics have brought as much heat to debates about human cognitive processes as *linguistic relativity* (Whorf, 1956), the idea that different grammars point their speakers to different evaluations of similar acts of observation (1956: 221). In perspectives on *categorisation*, defined here as a fundamental act of observation which underlies our understanding of the environment (Harnad, 1987), the main disagreement is between the relativist and the universalist approach. While the former advocates that categorisation behaviour varies depending on the structure of one’s native language, the latter defends the position that languages cannot influence categorical perception because categories are universal and innate. There is a growing body of experimental evidence that strongly suggests a greater perceptual refinement of categorical distinctions in speakers of a language where the given distinction is more prominently established (e.g., Casasanto, 2016; Everett, 2013, for recent overviews). As an example of this line of reasoning, one would expect to see faster category formation in English speakers than in Chinese speakers when the task is to sort change-of-state events based on phases (*the plane was running out of fuel* vs. *the plane ran out of fuel*). The rationale is that grammatical marking of different phases of change-of-state events is a feature prominent in the English, but not Chinese, aspect system. As a result, English speakers’ attention may be more strongly drawn to the distinction between the ongoing and completed phases of change-of-state events. This is due to frequent co-exposure to aspectual distinctions and perceptual contexts, where phases are the relevant diagnostic feature of category membership. Following Bylund & Athanasopoulos (2014), and Athanasopoulos et al. (2015), the underlying mechanism can be explained by the *attentional learning theory* (Smith & Samuelson, 2006). According to this theory, phase distinctions in this event type may be more accurate in English speakers because their relevant perceptual and linguistic associations are likely to be stronger as a result of encountering various event phases with their corresponding labels relatively more often. In contrast, attentional preferences of Chinese speakers to the initial phases would not be expected because co-occurrence of change-of-state events with linguistic marking of ongoingness in Chinese is very rare (Xiao & McEnery, 2004). Crosslinguistic differences in labelling patterns can in this way be expected to have an effect on discrimination patterns, and, through learned perception of category boundaries (Goldstone, 1994), predictively influence categorisation routines.

Importantly, consistent with both linguistic relativity and attentional learning theory, in English speakers one would predict to find increased attention allocation to phase distinctions, even in contexts where no overt linguistic labelling is necessary. This would suggest automatic activation of previously learned associations. Crosslinguistic differences are also expected to emerge in a silent categorisation task (i.e., where no language is overtly used) because if perceptual and linguistic information is frequently co-activated, it can provide sufficient training for automatic categorisation routines to occur (see e.g., Winawer et al., 2007, for colour discrimination). Nevertheless, language-specific effects are far from constant, and their contextual dependence can be tested by reducing the engagement of language in the categorisation process, for instance through verbal interference. According to the *label-feedback hypothesis* (Lupyan, 2012), ‘language produces *transient* modulation of ongoing perceptual and higher-level processing’ (2012: 4) including categorisation, and its contribution to categorical judgments can be regulated in both directions. It can be increased through verbal training (or linguistic priming) or decreased through verbal interference. The process of learning a second language provides verbal training of new kinds of visual-linguistic associations. Throughout this process, the dimensions relevant for L1-tuned categorisation become defocused, and learners become sensitised to new perceptual dimensions which the second language makes salient (Ellis, 2006). New combinations of perceptual and linguistic cues are likely to support new patterns of category organisation, which can result in changes of categorisation behaviour under the influence of the newly developing linguistic system. If L2-based and L1-based associations can facilitate category discrimination in similar ways, then it should also hold that L2-specific attentional biases in categorisation are similarly controllable, i.e., they should depend on the degree of L2 involvement. There are a few related points for specification. Framed within the label-feedback hypothesis, a context with overt verbal training in L2 would *up-regulate* the influence of relevant L2 labels and thus make discriminations involving them easier (e.g., Montero-Melis et al., 2016). A context with verbal distractors, in contrast, would *down-regulate* the impact of L2 labels and, in turn, draw less attention to previously learned associations between L2 labels, perceptual cues and the categorisation pattern this combination supports. Third, if we adopt a graded approach, the language of the distractors matters as well. It is possible that if verbal distractors themselves were in the L2, then L2-based categorisation would be weakened more than if verbal distractors were in a different language (Athanasopoulos et al., 2015). A context with categorisation performed in silence would serve as a suitable comparative baseline to detect whether L2 up- or down-regulation has any measurable impact on perceptual judgments. Some experimental demonstrations of these effects in perceptual learning contexts already exist.

**Perceptual learning studies with regulated access to verbal labels**

Research on the magnitude of influence language has on the cognitive dispositions of its speakers has recently grown through numerous captivating perceptual learning experiments. Studies testing possible effects of new or known verbal labels on category learning span various sensory domains, including vision (e.g., Lupyan et al., 2007; Kersten et al., 2010), sound (Dolscheid et al., 2013) and touch (Miller et al., 2018). For instance, participants in Lupyan et al. (2007) learned to categorise ‘aliens’ as approachable or to-be-avoided, based either on visual information alone or on visual information co-presented with a nonsense label. Labelling, even if not required to complete the task, facilitated category formation and lead to slightly more robust learning gains than visual information did on its own. In a different perceptual learning study with two language groups, Kersten et al. (2010) trained English and Spanish speakers to categorise ‘alien’ motion events based on manner versus path. English draws relatively more attention to the manner dimension in motion events (e.g., *a bug rolls away from the centre*) while Spanish typically foregrounds the path dimension (e.g., *a bug moves away from the centre rolling*). The English group was significantly more accurate in manner-based sorting, as predicted by linguistic relativity. Once the crosslinguistic contrasts were established, a separate experiment tested early and late Spanish-English bilinguals on the same task to see if L2-based categorisation would be more accurate in a second language context. Bilinguals exposed to English at an early age (<6y.) scored nearly as high on manner-based sorting as English natives, suggesting that immediate linguistic context can activate the mental links between L2 labels and the perceptual information it typically highlights. However, this study found no support for coactivation between L2 labels and the corresponding perceptual information in late (>6y.) bilinguals. A different result was found in Athanasopoulos et al. (2015), who observed that late German-English bilinguals (with an average L2 starting age of 11y.) can and do shift their categorisation preferences as a function of language context. One reason for such discrepant results is likely to be the different strength with which linguistic labels in the critical language were activated. While in the former study categorisation was performed in silence via button presses and the stimuli were co-presented with new (neither L1 nor L2-based) verbal labels, in the latter study the critical language was kept more active because bilinguals had to give overt linguistic responses in L1 or L2 to report their categorisation preference to the experimenter. The mosaic of these findings suggests that the processes of category formation and perceptual learning are adaptive, assisted by relevant L1 or L2 verbal labels as strongly as the current linguistic context activates them.

In another study of immediate relevance, Athanasopoulos and Albright (2016) presented a motion event categorisation test to speakers of English to examine the influence of verbal labels on perceptual learning. Participants were asked to categorise triads of video clips showing motion with low, medium, and high degrees of endpoint orientation (e.g., a man walking towards a car that is at a close, intermediate, and far distance). Their task was to choose whether the low or the high endpoint-biased clip is more similar to the model (i.e., the corresponding clip showing intermediate distance). In one condition, the correct answer was the low endpoint alternative, in the other condition it was the high one. No overt linguistic responses were required to complete the task, enabling the possibility of subvocal linguistic labelling. Assuming that participants engage in verbal labelling during silent categorisation, a performance advantage was expected in the low-endpoint condition. The rationale behind this prediction was that English, through grammatical aspect, habitually draws the attention of its speakers more to the initial phases rather than to the endpoints of unfolding motion events (von Stutterheim et al., 2012). The categorisation results were consistent with this prediction as participants scored significantly higher in the low endpoint condition. Learning either pattern took place at a comparable rate. This set of results provided an indication that increased accuracy in a low-endpoint condition may be linked to subvocal linguistic labelling, but the strength of this presumably language-induced facilitation remained unknown. In a separate experiment, the same study further explored this point using a verbal interference paradigm. If categorisation is verbally modulated, then the low-endpoint categorisation advantage was expected to disappear in a task with English number repetition alongside categorisation. This is exactly what the second experiment reported, together with the observation that learning took place in the high-endpoint condition but no longer in the English-typical condition. These findings brought important new insights into the selective advantage that verbal labelling in the speaker’s native language can contribute to the categorisation and learning of closely associated perceptual features. To substantiate a causal relationship, it would be informative to see whether comparable (or stronger) facilitation of learning and categorisation also emerges in a context where concordant verbal labelling is explicitly required, compared to when it is assumed to happen subvocally. To this end, the present study manipulates linguistic involvement and uses four versions of an event categorisation task, two in silence with possible access to native or L2 labels, the third with language involvement disrupted through verbal interference in L2, and the fourth with L2 labels up-regulated via explicit concordant verbal labelling during categorisation.

**Key concepts and experimental baseline**

The underlying premise this study builds on is the distinction between *action-oriented* and *result-oriented* categorisation of change-of-state events (e.g., *heavy wind closing the door*). *Action-oriented categorisation* is defined here as a preference to base event similarity judgments on the initial phase of an ongoing change-of-state event (*door closing*), while result-oriented categorisation is defined here as a preference to base such similarity judgments on the event’s final phase (*door closed*). English and Chinese speakers are expected to differ in their categorisation preferences because their native languages vary in the ways in which they make the perceptual dimension of an ongoing change-of-state event linguistically salient. English and Chinese both mark ongoingness grammatically (e.g., *Li* ***zai*** *xiang nei-jian shi* ‘Li **is** think**ing** about the matter’), but the English system may be better described as more action-oriented and Chinese as more result-oriented (Yong, 1997). This difference is attributable to the incompatibility of the Chinese ongoingness markers *zai* and *zhe* with change-of-state events, which renders constructions such as *\*Men zai guan/\*Men guan-zhe ‘*The door is closing’or *\*Fei zai dao jia/\*Fei dao-zhe jia* ‘Fei is arriving home’ highly infelicitous in Chinese (Klein et al., 2000; Xiao & McEnery, 2004). One line of support for this view comes from corpus data analyses examining the distribution of *–zhe* and *zai* across event types, and reporting that out of 238 instances of *–zhe,* not a single one was combined with change-of-state events expressed as achievements (Xiao & McEnery, 2004: 194), and out of 88 instances of *zai* only3 appeared with achievements (2004: 209). A brief diagnostic analysis was also conducted using the British National Corpus (BNC, 2007) to check the frequency of ongoingness marking for achievements in English. A set of achievement verbs that express change-of-state events was searched using the command ({kick} | {crash} | {extinguish} | {throw} | {squash} | {split} | {hang} | {drop} | {unhook} | {cut} | {catch} | {hit}). Out of the first 100 hits of a random sample, 14 were marked with a *V+ing* form as ongoing. These frequencies show that change-of-state events may not be most commonly expressed as ongoing in either language. Nonetheless, such combinations are relatively more common in English than they are in Chinese.

In this study, the term change-of-state event denotes biphasic situations with a short duration, which in linguistic form corresponds to the Vendlerian category of achievements, i.e., dynamic, telic, punctual predicates (Vendler, 1957). The expression of change-of-state events in Chinese exhibits slightly more tolerance to the progressive *zai* (e.g., w*omen zhengzai da-ying zhe-chang zhanzheng* ‘we are winning this war’) than to the durative *–zhe*, however, it holds that the compatibility of either marker with an achievement-type verb is very low. In contrast, the English equivalents interact much more readily arguably because the event phases selected for aspectual operations vary across the two language systems. Klein et al. (2000) reason that in English biphasic expressions (*the door is closing*), it is the source phase (i.e., *door not closed*) that gets selected for aspectual marking, while in Chinese it is the target phase (i.e., *door closed*). This distinction forms the conceptual frame of the present study, i.e., that English speakers’ perceptual judgments of biphasic change-of-state events can be expected to be more action-based and Chinese speakers’ categorisation more result-based due to the distributional constraints that make grammatical marking of ongoingness on the source phase more readily available in English than in Chinese.

To date there are two lines of experimental evidence (from verbalisation and categorisation) in support of the premise that English speakers are relatively more action-oriented in change-of-state events than Chinese speakers. Vanek and Selinker (2017) asked 42 Chinese and 42 English native speakers to verbally encode ten change-of-state events (such as *the boy is throwing away a frisbee*) under controlled input conditions. Participants watched 4-second animations and were tasked to describe each unfolding event in one sentence as soon as they recognised it. The study reported a substantially higher proportion of ongoingness marking in the verbalisations of change-of-state events in English (*M* = 56.43, *SD* = 29.3) than in Chinese (*M* = 21.43, *SD* = 24.38). Mixed-effects modelling was used to assess whether language group affiliation predicts ongoingness marking choices. With Group (English/Chinese) and Ongoingness marking (present/absent) used for the fixed effects, and Participant and Item used for the random effects, a significantly higher tendency to express change-of-state events as ongoing was reported for English speakers compared to Chinese speakers β = 5.89, SE = 1.51, Z = 3.89, p<.001 (a portion of the data for which participants gave their informed consent for public storage is available via [OSF](https://osf.io/gcy2u/) at <https://osf.io/6wvu3/> for an illustration). This analysis exhibited substantial crosslinguistic variation in how change-of-state events are expressed in English and Chinese. This difference motivated the assumption that a more frequent verbal expression of change-of-state events as ongoing gives rise to a relatively stronger preference in English speakers than Chinese speakers to establish similarity of unfolding change-of-state events (e.g., *throwing away a frisbee*) on the basis of their source phases (*frisbee being thrown*) rather than their target phases (*frisbee flying away*). Vanek and Selinker (2017) tested this assumption in an event categorisation task. To adequately capture the linguistic difference found in the verbalisation task, they modified the 4-second clips from the production experiment to create 2.5-second animation triads (available via the [IRIS](https://www.iris-database.org/iris/app/home/search?query=selinker) repository [www.iris-database.org](http://www.iris-database.org)). Each triad consisted of an equidistant clip (*frisbee being thrown* displayed for the same length of time than *frisbee flying away*), an action-oriented clip (*frisbee being thrown* displayed 4x longer than *frisbee flying away*), and a result-oriented clip (*frisbee being thrown* displayed 4x shorter than *frisbee flying away*). 21 Chinese and 21 English native speakers were asked to visually inspect the equidistant clip first. Then they watched the corresponding action-oriented and result-oriented clips played side by side underneath the equidistant clip. Their task was to decide which of the two clips they found most similar to the clip above them. In line with the patterns from the event description data, preference for action-oriented clips was operationalised as English-like categorisation and preference for result-oriented clips as Chinese-like categorisation. A mixed-effects regression model was fitted to the data to assess the contribution of language group to categorical choices. With Group (English/Chinese) used for the fixed effects, Categorisation (action-/result-oriented) as the outcome variable, and Participant and Item as random effect factors, the model showed that group membership was a significant predictor of categorical choices. English natives based their similarity judgments more frequently on the action-oriented source phases than Chinese natives β= −1.07, *SE* = 0.46, *Z* = −2.32, *p =* .02 (Vanek & Selinker, 2017: 235). The data are available via [OSF](https://osf.io/8snyw/) at <https://osf.io/8snyw/>. While target-phase choices in the English L1 group (*M* = 31.43, *SD* = 20.58) were significantly lower than chance (*p*<.001), in the Chinese L1 group (*M* = 46.67, *SD* = 32.19) they did not differ from chance performance (*p*>.05). The stronger inclination observed in the English speakers, compared with the Chinese speakers, to sort change-of-state events based on their source phases was interpreted as evidence that differences in grammatical aspect can significantly influence the assessment of perceptual features during event categorisation. In sum, the findings of Vanek and Selinker (2017) serve as the experimental baseline for the present study. The experimental baseline is important as it shows that different event categorisation patterns are attributable to language, rather than, for instance, to a generally higher saliency of boundaries in change-of-state events (Radvansky & Zacks, 2014). These findings provide empirical support to validate the assumption that English-like categorisation of change-of-state events is suitably characterised as action-oriented and Chinese-like categorisation as relatively more result-oriented.

**The present study**

The general motivation of this study is to test whether, and if so then how, subtle language manipulations can redirect second language learners’ cognitive behaviour. Two types of cognitive behaviour are measured, event categorisation accuracy and learning rate. Correct categorisation in this study depends on the learning condition; it is either based on the event’s source phase or on its target phase. Source-based (i.e., action-oriented or English-like) categorisation is taken to represents the less familiar type for Chinese learners of English. This assumption is built on the idea that ongoingness marking as the relevant diagnostic feature of category membership is an event feature more prominently established in English than in Chinese. A series of perceptual learning experiments is used with the aim to examine how strongly L2 English involvement affects L2-based categorisation. Three experiments were conducted under different learning conditions (in silence, with verbal interference, and with linguistic priming) to assess the role of L2 involvement during event categorisation. Different manipulations of L2 involvement follow the overarching logic that (a) explicit verbal labelling up-regulates linguistic involvement and facilitates concordant categorisation, (b) categorisation in silence is comparatively more difficult because verbalisation is optional and subvocal, and (c) categorisation with verbal interference is comparatively the most difficult because verbal distractors down-regulate linguistic involvement and hinder task performance.

To outline the series of experiments, Experiment 1 addresses the research question of whether L2-based categorisation is relatively more difficult than L1-based categorisation. The hypothesis is that learners’ accuracy rates will be lower in L2-based than in L1-based categorisation because, in the former, the links between linguistic material and the perceptual features they highlight are less well established. Significantly higher scores in L1-based than in L2-based categorisation would support this hypothesis, whereas no pronounced difference between the two categorisation types would disprove it. Experiment 2 addresses the research question of whether L2-based categorisation is less successful when performed with verbal interference than without it. The hypothesis is that verbal interference will reduce the availability of relevant L2 labels to aid L2-based categorisation, resulting in worse performance in a learning condition when verbal distraction is present than when it is not. Significantly lower accuracy scores in categorisation with verbal interference, compared to without verbal interference, would support this hypothesis, whereas no pronounced difference between the two categorisation types would disprove it. Experiment 3 addresses the research question of whether L2-based categorisation is more successful when performed with overt linguistic encoding than without it. The hypothesis is that overt verbalisation will up-regulate the availability of relevant L2 labels to assist L2-based categorisation, resulting in better performance in a learning condition where linguistic labels are overtly present than where they are not. Significantly higher accuracy scores in categorisation with, rather than without, overt verbalisation would support this hypothesis, no significant difference between these two categorisation types would disprove it.

**Experiment 1: Training action-oriented and result-oriented categorisation**

There are two research questions guiding this experiment. Is result-oriented (i.e., Chinese-like) categorisation relatively easier for Chinese learners of English than action-oriented (i.e., English-like) categorisation? Will block-to-block learning gains differ across conditions? A set of three hypotheses were formulated for analyses linked to these questions. If learners mentally encode events using verbal descriptions in their native language, result-oriented categorisation is expected to be initially easier, i.e., more accurate, than action-oriented categorisation. This is predicted by the linguistic relativity principle (Whorf, 1956, Casasanto, 2016; Everett, 2013), according to which categorisation is assisted by the learners’ native language because it makes the key perceptual dimension linguistically salient. Vanek and Selinker (2017) provide the experimental baseline here by showing that Chinese speakers tend not only to express but also to categorise ongoing change-of-state events as action-oriented significantly less often than English speakers. The null hypothesis is that categorisation is not linked to linguistic encoding in the native language (Jackendoff, 1990; Gleitman & Papafragou, 2005), in which case no significant difference is expected between initial accuracy scores in a result-oriented versus action-oriented condition. As for learning, block-to-block learning gains (Block 1 vs. 2-4) are not expected to differ across the two conditions (Kersten et al., 2010). The theory-building potential of this experiment is twofold. Framed within the attentional learning theory (Smith & Samuelson, 2006), a perceptual learning task with action-oriented versus result-oriented categorisation can show whether L2-based perceptual and linguistic associations are initially (Block 1) weaker than the L1-based ones. And if they are, comparisons of learning gains at later stages (blocks 1 vs. 2-4) can show whether the process of learning a new L2-based categorisation routine is or is not hindered by learners’ initial reliance on a non-L2-based system.

**Method**

**Participants**

Chinese learners of English who were enrolled in postgraduate studies in the UK at the time of testing formed the participant base. The inclusion criteria were normal or corrected-to-normal vision, no fluency in languages other than Chinese and English, and time spent in an English-speaking environment between 0-9 months (the participant details are available on the project website <https://osf.io/rxe6u/>). Further criteria included participation in an eligibility screening, which preceded the start of this study by 3 months. This screening involved 97 potentially eligible participants who (a) completed a general English proficiency test using the grammar section of the Oxford Placement Test 2 (Allan, 2004; the exclusion criterion was a score below 50/100), (b) filled in a linguistic background questionnaire (the exclusion criterion was a higher-than-beginner level/basic knowledge of a foreign language other than English), and (c) did an event description task in L2 English for stimulus validation purposes (the task was to verbalise stills showing 3 phases of 22 events out of which 50% was used in video format in this study). The exclusion criterion during stimulus validation was an inability to describe the main action in more than 2 events. Instructions before testing were limited to English only, participation was remunerated, and one experiment lasted 45 minutes on average. The target sample size for this experiment was to test 40 participants who met the inclusion criteria. At the stage of preregistration, this sample size was rationalised with the help of the G\*Power 3.1.9.2 software (Faul et al., 2007) available from [www.gpower.hhu.de](http://www.gpower.hhu.de) used to compute whether a medium or larger effect size (f ≥ 0.25) can be expected for this design (integral to the study preregistration through the Open Science Framework). Participants were randomly allocated to one of the equally sized predetermined training conditions (TCs). The event description pre-test showed that learners used the progressive aspect to express change-of-state events in their L2 English as ongoing (e.g., *the boy is hanging a hat on the hook*) in 17.35% of the cases. In 26.19% of the cases learners only marked the source phase of change-of-state events as ongoing (e.g., *the boy is jumping up to hang the hat*). In the remaining descriptions learners either expressed change-of-state events via the present perfect or the past simple (e.g., *the boy (has)put his hat on the hanger*).

**Materials**

Each participant was presented with four blocks of learning trials. One block consisted of 24 trials. One trial comprised three sequentially played videos depicting the same change-of-state event in three different phases: middle, source-oriented and target-oriented (examples available in the study design via <https://osf.io/rxe6u/>). The middle phase (always played first) was followed in 50% of trials by the source-phase and in the other 50% by the target phase. The order of trial presentation within a block was generated randomly for each participant. There was only one exception to full randomness, i.e., the first event shown in each block could not be the same as the last event shown in the preceding block in order to minimise possible mechanistic responses. The correct category in each block was the source phase for participants who were assigned to TC1, and the target phase for participants in TC2.

The sequenced videos consisted of a mid-phase clip played once (i.e., 62 frames played in 2.5 seconds cut out from the middle of a 100-frame/4sec. video, played at the rate of 25 fps) followed by a source-phase clip (i.e., the initial 62 frames from the corresponding 100-frame video) and a target-phase clip (i.e., the final 62 frames from the corresponding 100-frame video). The order of the source/target-phase clips was counterbalanced. The transition point (e.g., the moment of cutting off a branch) was played at 2.0 sec. in source-phase clips, at 1.25 sec. in mid-phase clips and at 0.5 sec. in target-phase clips. Each participant saw 24 video triads within one block. Following Athanasopoulos and Albright (2016), each triad was presented four times; twice in mid-source-target order and twice in mid-target-source order. There were 96 trials in total.

The tasks were programmed using Experiment Builder (v 1.10.1241, <https://www.sr-research.com/experiment-builder/>). Participants first saw the following written instructions: “You are going to see 3 short videos played in a sequence. Watch them very carefully. First you will see a Model video marked "M”, then a video marked “star”, and then a video marked “circle”. After the videos have all played, you must decide as fast as you can which video, “star” or “circle”, looks MOST like video “M”. Press the key marked “star” if you choose video “star”, or the key marked “circle” if you choose video “circle”. After your decision, a TICK will appear if your choice is correct, and a CROSS will appear if your choice is incorrect”.

After the instructions, the sequence for a single trial was as follows: blank screen (1000 ms), fixation cross in the middle of the screen (1000 ms), blank screen (1000 ms), mid-phase video marked “M” (2500 ms), blank screen (1000 ms), source/target-phase video marked “star” (2500 ms), blank screen (1000 ms), source/target-phase video marked “circle” (2500 ms), blank screen (1000 ms), instructions screen: “Choose which video is more similar to the model video “M”, “star” or “circle” (displayed until key press), feedback screen with a green tick or a red cross appearing over the answer and instructions underneath saying “Press SPACE to continue” (displayed until key press). Further details about the materials and the design can be found on the project’s website hosted by the Center for Open Science. The full collection of clips used in this study are shared with the research community through the above link to the project website. These clips are also available to the research community via the IRIS digital repository at <http://www.iris-database.org>.

**Procedure**

The experiment started with a key press once the participant confirmed they had understood the task. The categorisation rules were not revealed. The first group received training in English-like categorisation (TC1). In TC1, the first categorical treatment, the participants saw 3 videos in a sequence, a model video marked “M”, followed by a video marked “star” and a third video marked “circle”. After viewing the videos, the participant had to decide which video, “star” or “circle”, looks most like video “M”. The model video always showed the middle phase of the event. The “star” and the “circle” videos showed the source phase or the target phase of the same event. The manipulation lay in the training. In TC1, participants performed the task in silence, and a green tick appeared over the answer if they chose the source-phase clip, and a red cross appeared if they choose a target-phase clip.

The second group received training in Chinese-like categorisation (TC2). TC2 is the second categorical treatment in which the participant also saw 3 videos in a sequence, a model video marked “M”, followed by a video marked “star” and a third video marked “circle”. The task was the same, i.e., after viewing the videos to decide which video, “star” or “circle”, looks most like video “M”. The setup was identical to that in TC1, i.e., the model video was mid-phase, the “star” and the “circle” videos were either source or target phase. In TC2, the participants also performed the task in silence. The difference from TC1 was in the feedback. In TC2, a green tick appeared over the answer if participants chose the target phase, and a red cross appeared if they chose the source phase. After each block in either TC, participants were given a short break to rest (terminated via a key press) and were shown the number of correct responses they scored out of 24 for that block. No explicit information about the categorisation rule was given.

**Analysis plan**

The manipulated variables in this experiment are Training Condition (TC), varying between subjects, and Block, varying within subjects. The categorical variable of TC has two types, (TC1) training in English-like categorisation and (TC2) training in Chinese-like categorisation. Following previous categorisation studies with feedback (Kersten et al., 2010; Athanasopoulos & Albright, 2016), the measured dependent variable is the proportion of correct responses per participant in each block of trials. One block consists of 24 trials, i.e., 100% equals 24 correct answers. Each participant completed 4 blocks so that initial performance as well as improvement over time are quantifiable. Response correctness depended on condition. In TC1, a point was earned every time the participant chose the source phase video, and in TC2 the correct response was the target phase video.

*Non-preregistered analyses*

To test the initial set of hypotheses for Experiment 1, a generalized linear mixed-effects model was built using R (R Development Core Team, 2016) and *lme4* (Baayen et al*.*, 2008; Bates, Maechler, Bolker & Walker, 2014) to perform a series of analyses of the relationship between training condition and response accuracy. Training condition (English-like vs. Chinese-like) and Block (1-4) were entered as fixed effect factors, Participant and Item as random effect factors, and Response accuracy was the binary dependent variable. There were two steps in the modelling process. First, a subset of the data only from Block 1 was examined to answer the research question about initial response accuracy in English-like versus Chinese-like training conditions. The maximal model was planned (Barr et al., 2013) using the code glmer(CorrectResponse ~ 1 + TrainingCondition + (1 + TrainingCondition | Item) + (1 | Participant), family = “binomial”). The hypothesis was that training condition would be a significant predictor of initial response accuracy, namely that Chinese-like training would yield higher accuracy in Block 1 than English-like training. Second, the full dataset was examined to answer the research question about learning curves in English-like versus Chinese-like training conditions. This step was a more general comparison of accuracy rates in blocks 1-versus-2, 2-versus-3, and 3-versus-4, testing whether the overall rate of improvement differs as a function of training condition. For this reason, another (also maximal) model used the full dataset to test whether there is an effect of Training condition, and whether Training condition interacts with Block, using the code glmer(CorrectResponse ~ 1 + TrainingCondition \* Block + (1 + TrainingCondition \* Block | Item) + (1 + Block | Participant). Training condition was sum-coded and Block was forward-coded (following <https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/>). This coding system allowed the testing of whether block-to-block improvements systematically vary across learning conditions. The hypotheses were that the model will return a significant main effect of Training condition, which would indicate an advantage of Chinese-like training over English-like training, and that Block would also have significant main effects, indicating block-to-block improvements. No significant interactions between Training condition and Block were predicted because no systematic variation in learning gains was expected to depend on type of training. The annotated sequence of codes is available via <https://osf.io/6n3km/>. All stages of research reported in this article were guided by reproducibility and transparency principles (Nosek et al., 2015).

*Preregistered confirmatory analyses*

A mixed 2x4 ANOVA with Condition (TC1, TC2) as between-subject and Block (1-4) as within-subject factors was used following the pre-registered plan (available at <https://osf.io/sjp4t>). Using ANOVAs with categorical outcome variables (correct/incorrect) transformed to proportional data has been found to be problematic in the past (e.g., Jaeger, 2008; Johnson, 2009) because, among other reasons, ANOVAs are unable to account for random subject and item effects in a single-step analysis. To overcome this limit, the present article reports more informative mixed-effects models instead of the suboptimal pre-registered ANOVA-based analyses. A brief description of whether the findings of different analyses matched or not is provided in the results section for each experiment. In this article, more flexible mixed modelling takes precedence over less versatile ANOVAs. Detailed ANOVA results can be found in Supporting information online or via the project website at <https://osf.io/84zxm/>.

**Results** **of exploratory (non-preregistered) analyses**

First, to compare initial response accuracy between English-like versus Chinese-like training conditions (Table 1 shows the mean accuracy scores per block and training condition), a generalised linear mixed-effects model was built with Training condition (TC1 English-like vs. TC2 Chinese-like) and Block 1 as fixed effect factors, and Participant and Item as random effect factors. Response accuracy was the binary dependent variable. Participants with English-like training scored significantly lower than participants with Chinese-like training, β = -1.91, *SE* = .45, *Z* = -4.26, *p* <.001, which is in line with the related prediction.

**Table 1** Overview of the proportion of correct categorisation across training blocks in Experiment 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Training*** | ***Participants*** | ***Block 1***  ***M(SD)*** | ***Block 2***  ***M(SD)*** | ***Block 3***  ***M(SD)*** | ***Block 4***  ***M(SD)*** |
| English-like | 20 | .55 (.20) | .75 (.25) | .85 (.22) | .86 (.19) |
| Chinese-like | 20 | .84 (.19) | .92 (.19) | .96 (.12) | .98 (.06) |

Second, a more general comparison between TC1 and TC2 looked at whether the rate of progress from in blocks 1-to-4 also differs between the two training conditions. Training condition was found to be a significant contributor to response accuracy, β = -2.43, *SE* = .6, *Z* = -4.02, p<.001. A significant main effect of Training condition shows that participants with Chinese-like training (TC2) significantly outperformed participants with English-like training (TC1) (see Figure 1).



**Figure 1** Average proportion of correct event phase categorisations in training conditions 1 (English-like) and 2 (Chinese-like) across training blocks (error bars represent standard errors)

Follow-up pairwise comparisons were done to look more closely at block-to-block improvements. The first comparison examined the differences between the means of correct responses for Blocks 1 minus Block 2, the second comparison did the same for the differences between Block 2 minus Block 3 mean scores, and the third comparison for the differences between Block 3 minus Block 4 mean scores. The average score increase was significant in Block 2 compared to Block 1, β= -2.15, *SE* = .56, *Z* = -3.87, *p*<.001, but not in Block 3 compared to Block 2, β = -1.08, *SE* = .7, *Z* = -1.55, *p* = .12 or in Block 4 compared to Block 3, β = -.3, *SE* = .78, *Z* = -.381, *p* = .703. The significant main effect of Block indicates that participants significantly improved between the initial 2 blocks. There were no significant interactions between Training condition and block-to-block score increases (β = .76, SE = .66, *Z* = 1.14, *p* = .25 for Blocks 1 minus 2, β = -.3, *SE* = .7, *Z* = -.44, *p* = .66 for Blocks 2 minus 3, and β = .31, *SE* = .7, *Z* = .45, *p* = .66 for Block 4 minus 3), suggesting that, as predicted, learning rate was not influenced by training type. The main effects reported here matched the pattern of results from the preregistered ANOVAs <https://osf.io/84zxm/>. There was one mismatch, namely that, unlike mixed-effects modelling, the ANOVA returned a significant interaction between Training condition and Block. This discrepancy is likely due to the ANOVA analysis not taking into account random variation driven by participants and items.

**Discussion**

The categorisation test results of Experiment 1 confirm the prediction based on previous crosslinguistic research (Vanek & Selinker, 2017) that learners with Chinese L1 attend more strongly to the target phase than to the source phase in change-of-state events. Specifically, learners were initially more accurate in result-oriented categorisation than in action-oriented categorisation, suggesting that they noticed that target-biased clips can serve as a suitable basis to differentiate between event phases faster. Performance with Chinese-like training was found to be more successful throughout the task. These results provide new evidence for the principle of linguistic relativity in SLA (Bylund & Athanasopoulos, 2014). They demonstrate that the more prominently marked event feature in the learners’ native language has a significant influence on the assessment of perceptual features, even in contexts without overt verbalisation.

An improvement across blocks indicates success in learning the Chinese-like as well as the English-like pattern. Although the accuracy rate was overall higher in the Chinese-like condition throughout the four blocks, pairwise tests revealed that both types of training lead to a relatively fast pattern detection. This finding is consistent with previous supervised categorisation research (Athanasopoulos & Albright, 2016), and shows that attending more strongly to the perceptual dimension that is linguistically salient in the learners’ native language does not hinder the process of learning new patterns. An alternative explanation of the closely paralleled learning curves in the two conditions is the possibility that a more pronounced block-to-block improvement in TC2 was absent due to near-ceiling performance in this condition from early on. In sum, Experiment 1 tells us that, for Chinese learners of English, L2-based categorisation of change-of-state events is more difficult than L1-based categorisation. Experiment 2 continues the exploration of how strongly a second language can impact event categorisation by comparing learner performance in a categorisation task with reduced versus undisrupted access to L2 labels.

**Experiment 2: Training event categorisation with and without verbal interference**

To what extent is event categorisation verbally modulated? Will initial performance and improvement be hindered in a context where training is given concurrently with a verbal distractor? Initial performance was expected to be weaker in a verbal interference condition than in the silent condition because language availability is reduced and thus cannot assist with the categorisation task. This prediction is based on the label-feedback hypothesis (Lupyan, 2012), proposing that verbal interference down-regulates the influence of linguistic labels on categorisation. The null hypothesis is that categorisation depends purely on the perceptual dimensions of the stimulus, in which case initial and overall accuracy should not be affected by verbal interference. As for learning, improvement across blocks (1-4) in categorisation judgments was predicted to be smaller with concurrent verbal distraction than without it (Athanasopoulos & Albright, 2016).

**Participants**

In this experiment there was a new group of 20 participants trained with verbal interference. The inclusion criteria were identical to those for Experiment 1.

**Materials**

Experiment 2 used the same video materials as Experiment 1, with the addition of 20 pre-recorded number triplets read out loud in English. In TC3, the combination of video clips with the number triplet distractors was randomised. Unlike in TC1, the categorisation task in TC3 was performed while repeating numerical distractors. The information added to the instruction screen for TC3 was: “You will first hear three numbers, repeat them out loud. It is important that you keep repeating the three numbers until you make your decision”. The correct answer in TC3 was always the source-phase clip.

**Procedure**

In this categorical treatment, the participants first heard three random numbers between 11 and 99 that were read out in English. They were asked to repeat these numbers out loud throughout the trial, i.e., from just before the videos began up to making a choice between “star” or “circle” after they had viewed all three videos. The video order (“model” - “star” - “circle”) and the participants’ task were the same as in TC1 and TC2 in Experiment 1. After deciding which video, “star” or “circle”, looks most like video “M”, the number repetition could stop, and participants received feedback in the form of a green tick if the source phase was chosen, or a red cross if the target phase was chosen.

**Analysis plan**

*Non-preregistered analyses*

To test the second set of hypotheses for Experiment 2, a generalized linear mixed-effects model was built following the same logic as in Experiment 1. Training Condition (English-like categorisation with verbal interference, i.e., TC3 vs English-like silent categorisation, i.e., TC1) and Block (1-4) were entered as fixed effect factors, Participant and Item as random effect factors, and Response accuracy was the binary dependent variable. As in Experiment 1, the analyses took two steps. First, initial response accuracy (Block 1 scores only) was compared between the two training conditions. The prediction was that Training condition would be a significant predictor of response accuracy, namely that training with verbal interference would yield lower accuracy scores in Block 1 than training in silence would. Second, a more general comparison looked at the rate of learning gains between the two training conditions in blocks 1-4. A significant main effect of Training Condition was expected as a signal of a less difficult categorisation without verbal interference. Significant main effects of Block were also predicted as a signal of overall improvement across blocks. This step also planned to test whether there was an interaction between Condition and Block. An interaction was predicted because improvement was expected to be slower with verbal interference than without it.

*Preregistered confirmatory analyses*

A mixed 2x4 ANOVA with Condition (TC1, TC3) as between-subject and Block (1-4) as within-subject factors was also used following the pre-registration plan (available at <https://osf.io/sjp4t>). Just like for Experiment 1, more informative mixed-effects models are reported for Experiment 2 instead of suboptimal ANOVA-based analyses (available at <https://osf.io/84zxm/>). A brief comparison of whether different analyses yielded matching or mismatching results is provided in the results.

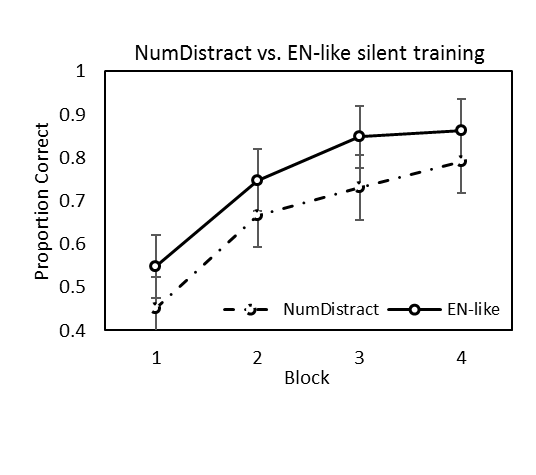
**Results** **of exploratory (non-preregistered) analyses**

To start with, initial response accuracy was compared from Block 1 between English-like training in silence versus English-like training with verbal interference (TC1 vs. TC3 respectively). Contrary to the related prediction, performance with verbal interference was not significantly worse than performance with silent training β= -.45, *SE* = .28, *Z* = -1.62, *p* = .1. Table 2 shows the mean accuracy scores in each block and training condition.

**Table 2** Overview of the proportion of correct categorisation across training blocks in Experiment 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Training*** | ***Participants*** | ***Block 1***  ***M(SD)*** | ***Block 2***  ***M(SD)*** | ***Block 3***  ***M(SD)*** | ***Block 4***  ***M(SD)*** |
| English-like | 20 | .55 (.20) | .75 (.25) | .85 (.22) | .86 (.19) |
| With distraction | 20 | .45 (.15) | .67 (.25) | .73 (.24) | .79 (.24) |

The second step was a more general comparison of the full dataset to see whether the overall rate of progress differs when learners are trained to categorise in silence versus with verbal interference. Training with verbal interference was less successful but not significantly less so than training in silence, β = -.21, *SE* = .45, *Z* = -.46, *p* = .64. Absence of a significant main effect of Training condition shows that participants who trained to categorise with a reduced access to L2 labels (TC3) were not significantly outperformed by participants who trained in silence (TC1) (see Figure 2).



**Fig****ure 2** Average proportion of correct event phase categorisations in training conditions 1 (English-like silent) and 3 (English-like with verbal interference) across training blocks (error bars represent standard errors)

Follow-up pairwise comparisons examined block-to-block improvements between the two training conditions. There was a significant score increase in Block 2 compared to Block 1, β = -1.41, *SE* = .38, *Z* = -3.7, *p*<.001, as well as from in Block 3 compared to Block 2, β = -1.18, *SE* = .33, *Z* = -3.57, *p*<.001, but not in Block 4 compared to Block 3, β = -.396, *SE* = .38, *Z* = -1.043, *p =* .297). These results indicate that participants exhibited significant improvements up until the penultimate versus the final block. Looking at interactions between Training condition and block-to-block score increases, learning progress was not significantly different in the verbal interference condition from that in the silent condition (β = .14, *SE* = .52, *Z* = .27, *p = .78* forBlocks 1 minus 2,β = .5, *SE* = .41, *Z* = 1.23, *p = .22* forBlocks 2 minus 3, andβ = -.22, *SE* = .43, *Z* = -.51, *p = .61* for Block 4 minus 3. This result suggests that verbal interference did not significantly slow down the overall rate of learning. The corresponding preregistered ANOVA test (<https://osf.io/84zxm/>) matched this pattern of results.

**Discussion**

The results of Experiment 2 were inconsistent with the prediction that performance with silent training would be significantly better than with number distractors. The trend of lower accuracy scores in a verbal interference condition across blocks is in line with the label-feedback hypothesis (Lupyan, 2012), however, analyses showed that event differentiation based on source phases is not significantly more difficult in a context where the possibility of linguistic labelling is down-regulated. The main effect of Block showed that successful learning of the English-like pattern did occur even when the option of silent verbalisation was reduced. The significant accuracy increase early in the learning process (from Block 1 to Block 2) is particularly notable despite consistent and robust verbal distraction.

Unlike in previous research with a concurrent nonverbal task during categorisation training (Athanasopoulos & Albright, 2016), learning rate in this experiment was not affected by verbal distraction. One explanation of this discrepancy might be that in the present study the videos were more homogeneous (with the protagonist, the direction of movement and the background kept identical across clips), thus making the critical pattern more consistent across items and possibly more easily detectable in comparison. Another possible explanation is that learners did not use verbal labelling during silent categorisation. Thus, minimising linguistic involvement did not make a difference to either initial and overall accuracy or to the rate of learning gains. It is reasonable to assume little activation of the relevant verbal labels in this specific group of L2 users, considering that in the pre-test they only expressed 17.35% of change-of-state events as ongoing. For an informative comparison, Experiment 3 was designed to see how relevant L2 labels influence change-of-state event categorisation when they are fully activated by overt use.

**Experiment 3: Training event categorisation with and without overt verbalisation**

Will categorisation performance be relatively better and improvement relatively faster in a context where training is given concurrently with overt linguistic encoding of the key perceptual feature? The hypothesis was that consistent pairing of overt linguistic labels in English with the corresponding clips will facilitate English-like categorisation, with initial performance and overall improvement expected to exceed the scores of participants trained using silent English-like categorisation. This prediction is based on the label-feedback hypothesis (Lupyan, 2012) according to which overt verbalisation can up-regulate the contribution of linguistic labels (i.e., make the key perceptual feature more prominent), thus assisting the learning of novel categories (Lupyan et al., 2007). The null hypothesis is that L2 users are not relying on the provided L2 labels to assist categorisation, in which case there will be no effect of overt verbalisation on performance and improvement in comparison with silent categorisation.

**Participants**

For Experiment 3, a new group of 20 participants was recruited and trained to categorise with overt verbalisation. The inclusion criteria were the same as those for Experiment 1.

**Materials**

In Experiment 3, the same video materials were used as in Experiment 1, with the addition of 24 pre-recorded model sentences read out loud in English for TC4. The unique feature in TC4 was that categorisation was preceded by overt verbalisation of the corresponding event, i.e., participants were asked to repeat a model sentence (see an example and details in Procedure). The information added to the instruction screen for TC4 was: “You will first hear a sentence, repeat this sentence out loud”. The correct answer in TC4 was always the source-phase clip.

**Procedure**

This experiment involved a group of 20 participants trained with model sentence repetition (TC4). In TC4, the fourth categorical treatment of this study, participants first heard a sentence, which they were asked to repeat out loud. Each model/prime sentence was a verbal expression of the event shown in the upcoming video. The model sentences were constructed to contain one achievement verb in *V+ing* form as well as the event’s result (e.g., *The boy is throwing away a frisbee*). When the participant misformulated the *V+ing* form or omitted the result, the experimenter asked for an accurate second attempt with both components. Then, the three videos were played in the usual sequence (“model” -> “star” -> “circle”). The task was the same as in TC1-3. After deciding which video, “star” or “circle”, looked most like video “M”, a green tick appeared if the source phase was chosen, or a red cross appeared if the target phase was chosen.

**Analysis Plan**

*Non-preregistered analyses*

To test the third set of hypotheses for Experiment 3, a generalized linear mixed-effects model was built following the same logic as in the previous two experiment. Training condition (categorisation following overt verbalisation, i.e., TC4 vs. English-like silent categorisation, i.e., TC1) and Block (1-4) were the fixed effects factors, Participant and Item were the random effect factors, and Response accuracy was the binary dependent variable. The analyses took two steps. In the first step, response accuracy was compared between Block 1 scores for training with overt verbalisation and those for training in silence. The prediction would be confirmed if, for Block 1, training with overt verbalisation yielded significantly higher accuracy scores compared to training in silence. In the second step, the overall rate of learning gains was compared between the two training conditions. A main effect of Training condition was expected to show an overall advantage when the relevant linguistic labels were overtly verbalised. Main effects of Block were predicted to show improvements across individual blocks, and a significant interaction between Training condition and Block was predicted to show faster improvement across blocks in training with overt verbalisation.

*Preregistered confirmatory analyses*

A mixed 2x4 ANOVA with Condition (TC1, TC4) as between-subject and Block (1-4) as within-subject factors was also run following the pre-registered plan (available at <https://osf.io/sjp4t>). Just like for the previous experiments, more powerful mixed-effects models are reported for Experiment 3 instead of the suboptimal ANOVA-based analyses (available at <https://osf.io/84zxm/>). A brief comparison of whether different analyses yielded matching or mismatching outcomes is provided in Results.

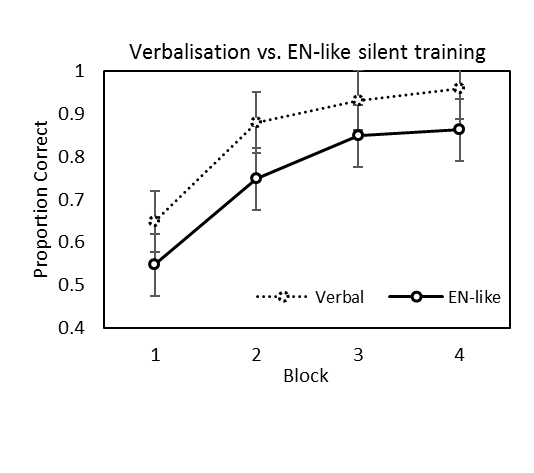
**Results of exploratory (non-preregistered) analyses**

Comparing initial accuracy in Block 1 only, performance with model sentence repetition was not significantly better than performance with silent training, β = .496, *SE* = .31, *Z* = -1.62, *p* = .11. This result goes against the related prediction that accuracy in the initial block of training with second language up-regulated would be significantly better than with silent training. Table 3 shows the mean accuracy scores in each block and training condition.

**Table 3** Overview of the proportion of correct categorisation across training blocks in Experiment 3

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Training*** | ***Participants*** | ***Block 1***  ***M(SD)*** | ***Block 2***  ***M(SD)*** | ***Block 3***  ***M(SD)*** | ***Block 4***  ***M(SD)*** |
| English-like | 20 | .55 (.20) | .75 (.25) | .85 (.22) | .86 (.19) |
| With verbalisation | 20 | .64 (.20) | .88 (.16) | .93 (.13) | .96 (.08) |

The second step of the analysis was a more general comparison to see whether the overall rate of progress is better in English-like training with model sentence repetition than in English-like silent training. The mixed-effects model returned a significant difference in categorisation accuracy between training with overt verbalisation and silent training, β = 1.09, *SE* = .46, *Z* = 2.35, *p* = .019. A significant main effect of Training condition confirms the related prediction and shows that participants who trained to categorise with L2 labels up-regulated (TC4) significantly outperformed participants who trained in silence (TC1) (see Figure 3).



**Figure 3** Average proportion of correct event phase categorisations in training conditions 1 (English-like silent) and 4 (English-like with model sentence repetition) across training blocks (error bars represent standard errors)

Next, follow-up pairwise comparisons were used to examine block-to-block improvements between the two training conditions (TC4 vs. TC1). A significant score increase was found in Block 2 compared to Block 1, β = -1.39, *SE* = .34, *Z* = -4.06, *p* <.001, as well as from in Block 3 compared to Block 2, β = -1.33, *SE* = .39, *Z* = -3.41, *p*<.001, but not in Block 4 compared to Block 3, β = -.018, *SE* = .35, *Z* = -.05, *p =* .96). These results signal substantial block-to-block improvements, which plateaued at the penultimate-to-final block stage. No interactions between Training condition and block-to-block score increases were significant (β = -.63, *SE* = .55, *Z* = -1.33, *p* = .18forBlocks 1 minus 2,β = -.02, *SE* = .55, *Z* = -.03, *p* = .98 for Blocks 2 minus 3, andβ = -.38, *SE* = .5, *Z* = -.76, *p* = .45for Block 4 minus 3). This result suggests that, contrary to the related prediction, overt verbalisation of the key perceptual feature did not significantly speed up L2-based categorisation. The corresponding preregistered ANOVA tests (<https://osf.io/84zxm/>) matched this pattern of results, including a significant main effect of Training condition.

**Discussion**

One prediction of Experiment 3 was confirmed by the results, two were not. First, training with overt verbalisation was found to be numerically more successful than training in silence, but the difference for the first block did not reach statistical significance as predicted. However, later on, i.e., after Block 1, verbalisation of the key perceptual feature provided the expected boost and led to a significantly higher categorisation accuracy than training in silence did. This effect of verbalisation suggests that attention to event distinctions based on their source phases is stronger in a context where linguistic responses are necessary than in a context where they are optional. Boosted performance with overt language use is interpreted as support for the label-feedback hypothesis (Lupyan et al., 2007; Lupyan, 2012), according to which overt linguistic labelling up-regulates the prominence of verbalised perceptual features and aids related category formation.

Out of all three English-like training conditions, it was only training with verbalisation that matched the average accuracy level achieved with Chinese-like training (in Block 3). Even though a significant improvement was observed across blocks in both training conditions, explicit verbalisation did not facilitate learning speed that would significantly exceed that observed in silent categorisation. This result refutes the related hypothesis and indicates that even though overt verbalisation may bring some advantage during event classification, it does not strongly influence the rate of perceptual learning. A possible reason for no rate advantage when L2 labels were up-regulated could be that for this group of late learners, as in Kersten et al. (2010), the linguistic material did not activate mental links to perceptual information with sufficient strength to have an online effect on the speed of category formation.

**General discussion**

The present study examined the extent to which a second language facilitates learning a new way of event categorisation. To this end, perceptual judgments with various involvement of the L2 served as the testbed for comparing the influence of an L2-based labelling pattern on new category formation. Three sets of findings emerged. First, Experiment 1 showed that Chinese learners of English rely more strongly on L1-based result-oriented event categorisation than on L2-based action-oriented event categorisation. Learners not only noticed that result-biased clips can be a suitable foundation for distinguishing event phases faster, but L1-based training also proved more effective than L2-based training throughout the learning task. These results align with the linguistic relativity approach as they clearly signal that, in a context without overt verbalisation, L1-based event categories are more helpful for perceptual feature assessment than L2-based event categories. Second, Experiment 2 compared the impact of two types of L2-based training, where one group trained in silence and the other group with verbal interference. Response accuracy was numerically lower in the verbal interference condition, but the difference was not statistically significant. This absence of a predicted effect of verbal interference suggests that participants may have judged videos predominantly on the basis of their visual properties without resorting to a relevant verbal strategy to aid silent categorisation. Minimising the involvement of language did not significantly increase the difficulty of correct event differentiation nor did it impact the learning rate. One possible explanation is that L2 labels for event phase distinctions, such as *is closing* versus *has closed*, were not activated or retrieved to assist categorisation performance. Third, in Experiment 3 the comparison of L2-based training with and without overt language production showed that response accuracy was significantly higher in a context that maximised the involvement of relevant L2 labels. However, again, the learning rate remained unaffected. These results suggest that overt linguistic labelling in the L2 facilitates perceptual judgments when an L2-based category is formed, but its impact on the speed of learning a new categorisation pattern is negligible. Overall, the findings highlight that the degree of second language involvement can influence L2-based categorisation behaviour but not the learning rate. Importantly, they show that up-regulating specific aspectual properties of the second language system steers attention towards the corresponding event features and modulates L2 learners’ perceptual judgments.

A growing body of experimental studies shows that verbal labelling in one’s native language, including lexical (Smith & Samuelson, 2006) as well as grammatical labelling (Lucy & Gaskins, 2001), benefits category recognition and learning. When verbal labels in the speaker’s first language correlate with perceptual regularities, these correlates enhance L1-based discrimination between categories and facilitate perceptual judgments about them (Goldstone & Hendrickson, 2009). Consequently, crosslinguistic differences in how strongly specific labelling routines correlate with perceptual regularities give rise to linguistic relativity effects in non-verbal contexts (e.g., Casasanto, 2016; Everett, 2013 for an overview). This study brings new evidence for linguistic relativity by showing that Chinese learners of English categorise change-of-state events less accurately when category membership relies on grammatical contrast in phase marking, a feature prominent in the English but not in the Chinese aspect system. Framed within the attentional learning theory (Smith & Samuelson, 2006), higher accuracy in result-oriented event phase distinctions of Chinese learners of English is attributable to relatively stronger L1-based perceptual and linguistic associations.

It is important to view linguistic influence on categorisation patterns as a scalar phenomenon because perceptual judgments are not automatically based on features of the visual world that the native language makes salient (Gennari et al., 2002; Papafragou & Selimis, 2010). The strength of effects that linguistic labels have on categorisation performance is context-induced and flexible rather than permanent. In contexts where L2 labels can serve as a more reliable diagnostic tool than L1 labels to establish category membership, such as in Experiment 3, perceptual judgments appear to dynamically adapt from L1-specific to L2-tuned categorisation. Findings that suggest this fit well with related research which shows that recent L2 priming affects event similarity assessment (Montero-Melis et al., 2016). They also extend previous inquiry by documenting the extent to which perceptual learning is influenced when L2 labelling is required relative to when it is optional (Athanasopoulos & Albright, 2016). A direct comparison where these two training contexts are the only manipulated variable is important because it demonstrates that explicit verbalisation (or up-regulation, Lupyan, 2012) of corresponding L2 cues changed L1-based perceptual judgments to L2-based with overall more success than optional verbalisation in silence. In other words, increasing second language involvement strengthens the structural basis, which provides learners with online feedback between perceptual features and the corresponding L2 labels. This up-regulation effect fits well with the theories of attentional learning (e.g., Smith & Samuelson, 2006), namely with the mechanism that associations are built when learners frequently encounter linguistic material co-occurring with related perceptual information. With a closer link to the role of the L2 (following Bylund & Athanasopoulos, 2014), these results show for the first time that when aspectual information is highlighted via overt use of the L2 grammar, L2 learners gradually shift their evaluations of perceived event features. The fact that learners became cognitively more committed to L2-based categorisation through overt linguistic encoding is interpreted as a signal that the ways in which we evaluate events is substantially modulated by the grammatical properties of the currently used language.

Down-regulating the impact of the L2 did not work quite as expected. When the possibility of recruiting linguistic labels to aid the categorisation process was minimised through verbal interference, category discrimination became more difficult, but not significantly more so than in the silent condition. This absence of a verbal interference effect on performance accuracy shows that, in the silent training condition, retrieval of relevant grammatical information in the second language was not strong enough to influence perceptual judgments in an L2-based fashion. Framed within theories of attentional learning (e.g., Smith & Samuelson, 2006), the link between Chinese learners’ attentional biases and the grammatical encoding of unfolding change-of-state events as ongoing was not (yet) as well established as in English native speakers (Vanek & Selinker, 2017). An alternative account for the lack of structural modulations when access to the L2 verbal system is disrupted is that learners did not recruit the critical linguistic structures to aid their categorisation performance in the silent condition. This could be because the key grammatical constructions go beyond single labels. Traditionally, the effect of verbal interference to down-regulate the retrieval of labels has been tested with perceptual features, whose linguistic realisation is comparably simpler, through single words such as colours (e.g., Winawer et al., 2007) and single-word event labels to encode path versus Manner, such as *ascending* versus *sliding* (e.g., Trueswell & Papafragou, 2010). Categorising phases of change-of-state events might be different in this respect because the relevant linguistic features are essentially more complex, e.g. *throwing away* versus *threw away*. With increased complexity, perceptual modulations under the influence of multi-word constructions might work differently than with single labels, especially when the relevant labelling is in a second language.

Measuring the success of L2-based categorisation block by block in this study allowed for the tracing of how strong the boost of L2 labels is while new perceptual distinctions are being learned. Analyses showed a significant advantage of categorisation training with overt verbalisation over training in silence. With overt verbal labels in L2, learners did become overall more sensitised to a new L2-based perceptual dimension, but it remains a future challenge to verify whether the advantage to retune L1-based categorisation observed here also holds across other participant pools, item configurations and grammatical aspect systems. Another important point concerns the control of L2-specific attentional biases through the type of verbal distraction. Previous research (Athanasopoulos et al., 2015) suggests that when access to L2 is disrupted, bilinguals tend to shift their categorisation preferences more strongly to the patterns associated with the L1, and the same was shown to hold vice versa. In this study, the distractor language was the L2, which enabled the measurement of how malleable L2-based categorisation is with and without L2 involvement. In an extended design, manipulating the language of distraction would be a suitable alternative to assess the magnitude of bilingual speakers’ reliance on the L2 in their perceptual judgments.

**Limitations and future research**

A number of thought-provoking questions remain open. For instance, how deeply does event perception change when one learns a new language? In other words, on which level(s) do aspectual cues in the L2 activate the visual properties of events they refer to? Is their influence limited to conscious behavioural responses, or do they also have a comparable impact on the L2 learners’ automatic perceptual processing detectable via brain responses? And if so, do different types of perceptual training modify learners’ brain activity in predictable ways? Crosslinguistic research shows that grammatical aspect can influence brain activity during event perception in language-specific ways (Flecken et al. 2015). Related research testing the effects of L2 feedback (e.g., on risk taking, Gao et al., 2015) signals that a greater difference in behavioural responses can be expected to correlate with a greater difference in brain responses. An extension of the present paradigm to the neurophysiological domain through the combination of regulated L2 involvement, perceptual learning, and event-related brain potential measurement, could provide a helpful tool to shed light on the transformative power of L2-based training on automatic perceptual as well as conscious decision-based processing.

Another emerging question concerns the extent to which variation in the obtained categorisation patterns was driven mainly by the crosslinguistic difference in marking events as ongoing, or whether differences in marking events as completed also played a role. Chinese encodes event completion structurally via the perfective *le* and via resultative verb constructions (RVCs) such as *Qi-che* ***zhuang-dao*** *-le fangzi* ‘The car (**hit-break**)/knocked down the house’ (Klein et al., 2000: 724); English does not. Given that completion marking is more prominently established in Chinese, it may be a more relevant diagnostic feature for event categorisation for Chinese speakers than for English speakers. Some support for the idea that Chinese may be relatively more result-focused rather than just less ongoingness-focused comes from production data in Vanek and Selinker (2017) available at <https://osf.io/6wvu3/>, showing that 51.15% out of 260 change-of-state event descriptions were explicitly marked as completed (via *le* and/or RVCs) by Chinese speakers compared to only 6.76% out of 322 marked as completed (via past simple) by English speakers. To assess more precisely how much ongoingness marking in English versus completion marking in Chinese contribute to event category formation in Chinese learners of English, future perceptual learning studies would benefit from employing a fully crossed 2x3 design with the three learning conditions (silent, linguistically primed, with interference) tested in both source-based and target-based categorisation tasks. A further suitable extension would be adding English and Chinese native speakers to the present design. A perceptual learning experiment with native speaker groups doing the same tasks with the same stimuli would serve as a more solid comparative baseline than a similar categorisation task (Vanek & Selinker, 2017) with only a subset of the current stimuli. Keeping the paradigm and the stimuli fully identical would help to establish more firmly whether source-based categorisation is relatively less difficult for English speakers.

A separate question is why verbal interference did not cause categorisation impairments (Lupyan, 2009) as predicted. There are several possible accounts. One explanation is that number repetition in the dual-task group did not down-regulate the assumed covert contribution of linguistic labels from the very first block because sufficiently strong links between L2 representations and conceptual representations in the tested learners had not (yet) been formed. Future designs could profit from including a wider range of L2 users, including more advanced individuals, to test whether mobilisation of L2-specific linguistic resources during silent event categorisation changes as a function of L2 proficiency. In addition, future studies may find it an analytical improvement to maximise parsimony by building just one mixed-effects model with appropriate coding that would answer questions about both accuracy and learning rate differences. Returning to verbal interference, an alternative account is that no effect was observed because distractor repetition and label-modulated event categorisation operate at different levels of language processing. If we assume that visual-linguistic associations pertain to the level of conceptualisation but repeating numbers is limited to the level of articulation, then it is unlikely that access to language-related concepts would be blocked by articulating numbers because these processes can take place in parallel. To address this issue, further studies could find it helpful to manipulate language-related distractors on the same level of processing where perceptual and linguistic cues integrate to form language-related concepts. This could be achieved, for instance, by measuring learners’ behavioural responses in two event categorisation tasks, one with the involvement of an L2-related distractor, such as a picture of a source-phase mismatched event, and the other without it. Manipulating attentional biases to specific event phases on the level of conceptualisation will be useful in future perceptual learning studies because it will allow us to better understand when and how strongly the development of new processing routines is L2-related.

**Conclusion**

In this study, a novel combination of methods allowed for tracking of L2 users’ perceptual judgments in four types of categorisation training. This approach has proven to be useful to connect theories of attentional and perceptual learning (Goldstone, 1998; Goldstone & Hendrickson, 2010; Smith & Samuelson, 2006) with a second language context (as suggested in Bylund & Athanasopoulos, 2014 and Athanasopoulos et al., 2015). The present findings provide evidence that attentional biases in event categorisation are controllable through a regulated involvement of L2 labels that make the relevant perceptual dimension salient. Changes in categorisation behaviour under the influence of the L2 (Ellis, 2006) demonstrated that learners readily integrate new associations between L2 labels and related visual properties into their perceptual judgments, depending on how accessible the relevant L2 labels are in a given learning context. However, the findings also show that while category recognition was facilitated in predictable ways, learning rate continued to improve comparably across conditions, regardless of how accessible the second language was. This dissociation between L2 label accessibility and learning rate differs from earlier findings in research that manipulated L1 access during classification (Athanasopoulos & Albright, 2016), and suggests that gradual increases in accuracy were facilitated by cues other than L2 labels. One plausible explanation for this difference is that L2-based associations between percepts and their corresponding L2 labels may make category discriminations involving them easier, but they are not strong enough to aid learning to the same extent that L1-based associations are. These insights are important in two ways. First, they inform attentional and perceptual learning theories about the nature of grammatical cue integration when a new L2-based event category is formed. Second, and crucially, they help to enhance understanding of the changes in nonverbal behaviour under the influence of a second language.

Behavioral findings from previous research show that there is a selective advantage of verbal labelling when speakers learn categories on the basis of those perceptual features that their native language makes salient (Kersten et al., 2010). Later work testing the cognitive performance of bilingual speakers shows that category formation can also be facilitated when the verbal labels congruent with the key perceptual features are in the speaker’s second language rather than the first (Athanasopoulos et al., 2015, Montero-Melis et al., 2016). This study further extends previous work on how formal aspects of the L2 can help refine perceptual category distinctions. With systematic regulation of L2 access, this is the first time that crosslinguistic differences in grammatical aspect have been investigated in a set of four perceptual learning contexts. The findings provide evidence for a graded impact of L2 grammatical aspect on the success of category formation, found to be highest with overt verbalisation in the L2, followed by categorisation in silence, and the lowest but not significantly reduced with verbal interference in the L2. These findings underline that L1-tuned categorisation can be rapidly defocussed and flexibly overdriven by L2-based categories. They also highlight how robustly L2 cues support new patterns of category organisation by showing that the more second language labels are involved in a categorisation task, the stronger L2-specific attentional biases they induce.

**References**

Allan, D. (2004). *Oxford placement test 2: Test pack*. Oxford: Oxford University Press.

Athanasopoulos, P., Bylund, E., Montero-Melis, G., Damjanovic, L., Schartner, A., Kibbe, A., Riches, N. & Thierry, G. (2015). Two languages, two minds: flexible cognitive processing driven by language of operation. *Psychological Science, 26 (4),* 518 – 526. <https://doi.org/10.1177/0956797614567509>

Athanasopoulos, P., & Albright, D. (2016). A perceptual learning approach to the Whorfian hypothesis: Supervised classification of motion. *Language Learning, 66(3),* 666–689. <https://doi.org/10.1111/lang.12180>

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68(3)*, 255-278. <https://doi.org/10.1016/j.jml.2012.11.001>

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version, 1(7),* 1-23.

Bylund, E., & Athanasopoulos, P. (2014). Linguistic relativity in SLA: Toward a new research program. *Language Learning, 64*, 952–985. <https://doi.org/10.1111/lang.12080>

Bylund, E., Athanasopoulos, P., & Oostendorp, M. (2013). Motion event cognition and grammatical aspect: evidence from Afrikaans. *Linguistics, 51,* 929 – 955. <https://doi.org/10.1515/ling-2013-0033>

Casasanto, D. (2016). *Linguistic relativity.* In N. Riemer (ed.) *The Routledge handbook of semantics*, pp. 158-174, New York: Routledge.

Dolscheid, S., Shayan, S., Majid, A., & Casasanto, D. (2013). The thickness of musical pitch: Psychophysical evidence for linguistic relativity. *Psychological Science, 24(5)*, 613-621. <https://doi.org/10.1177/0956797612457374>

Ellis, N. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics, 27(2),* 164–194. <https://doi.org/10.1093/applin/aml015>

Everett, C. (2013). *Linguistic relativity: Evidence across languages and cognitive domains*. Berlin: de Gruyter.

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G\*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods, 39*, 175–191. <https://doi.org/10.3758/BF03193146>

Filipović, L. (2011). Speaking and remembering in one or two languages: bilingual vs. monolingual lexicalization and memory for motion events. *International Journal of Bilingualism, 15(4)*, 466-485. [https://doi.org/10.1177/1367006911403062](https://doi.org/10.1177%2F1367006911403062)

Flecken, M., Athanasopoulos, P., Kuipers, J. R., & Thierry, G. (2015). On the road to somewhere: Brain potentials reflect language effects on motion event perception. *Cognition, 141*, 41-51. <https://doi.org/10.1016/j.cognition.2015.04.006>

Gao, S., Zika, O., Rogers, R. D., & Thierry, G. (2015). Second language feedback abolishes the "hot hand" effect during even-probability gambling. *Journal of Neuroscience, 35*, 5983–5989. <https://doi.org/10.1523/JNEUROSCI.3622-14.2015>

Gennari, S., Sloman, S., Malt, B., & Fitch, W. (2002). Motion events in language and cognition. *Cognition, 83(1),* 49–79. <https://doi.org/10.1016/S0010-0277(01)00166-4>

Gleitman, L., & Papafragou, A. (2005). Language and thought. In K. Holyoak & R. Morrison (eds.), *Cambridge handbook of thinking and reasoning* (pp. 633–661). Cambridge: Cambridge University Press.

Goldstone, L. (1998). Perceptual learning. *Annual Review of Psychology, 49*, 585–612. <https://doi.org/10.1146/annurev.psych.49.1.585>

Goldstone, L., & Hendrickson, A. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 1(1),* 69-78. <https://doi.org/10.1002/wcs.26>

Harnad, S. (1987). *Categorical perception: The groundwork of cognition*. Cambridge, UK: Cambridge University Press.

Jackendoff, R. (1990). Semantic structures. Cambridge, MA: MIT Press.

Jaeger, F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59(4),* 434–446. <https://doi.org/10.1016/j.jml.2007.11.007>

Johnson, D. E. (2009). Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass, 3(1)*, 359–383. <https://doi.org/10.1111/j.1749-818X.2008.00108.x>

Kersten, A., Meissner, C., Lechuga, J., Schwartz, B., Albrechtsen, J., & Iglesias, A. (2010). English speakers attend more strongly than Spanish speakers to manner of motion when classifying novel objects and events. *Journal of Experimental Psychology: General, 139*, 638–653. <https://doi.org/10.1037/a0020507>

Klein, W., Li, P., & Hendriks, H. (2000). Aspect and assertion in Mandarin Chinese. *Natural Language & Linguistic Theory, 18(4)*, 723–770. <https://doi.org/10.1023/A:1006411825993>

Lucy, J., & Gaskins, S. (2001). Grammatical categories and the development of classification preferences: A comparative approach. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 257–283). Cambridge: Cambridge University Press.

Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences of the United States of America, 110*, 14196–14201. <https://doi.org/10.1073/pnas.1303312110>

Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology, 3*, 1–13. <https://doi.org/10.3389/fpsyg.2012.00054>

Lupyan, G. (2009). Extracommunicative functions of language: Verbal interference causes selective categorization impairments. *Psychonomic Bulletin & Review, 16(4),* 711-718. <https://doi.org/10.3758/PBR.16.4.711>

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: redundant labels facilitate learning of novel categories. *Psychological Science, 18*, 1077–1083. <https://doi.org/10.1111/j.1467-9280.2007.02028.x>

Marsden, E., Mackey A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS Repository of Instruments for Research into Second Languages* (pp. 1-21). New York: Routledge.

Miller, T. M., Schmidt, T. T., Blankenburg, F., & Pulvermüller, F. (2018). Verbal labels facilitate tactile perception. *Cognition, 171*, 172–179. <https://doi.org/10.1016/j.cognition.2017.10.010>

Montero-Melis, G., Jaeger, T. F., & Bylund, E. (2016). Thinking is modulated by recent linguistic experience: Second language priming affects perceived event similarity*. Language Learning, 66(3)*, 636–665. <https://doi.org/10.1111/lang.12172>

Nosek, B. et al. (2015). Promoting an open research culture. *Science, 348*, 1422–1425. <https://doi.org/10.1126/science.aab2374>

Papafragou, A., & Selimis, S. (2010). Event categorisation and language: A cross-linguistic study of motion. *Language and Cognitive Processes, 25*, 224–260. <https://doi.org/10.1080/01690960903017000>

R development core team (2016). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, <http://www.R-project.org>

Radvansky, G. A., & Zacks, J. M. (2014). *Event cognition*. Oxford: Oxford University Press.

Smith, L., & Samuelson, L. (2006). An attentional learning account of the shape bias: Reply to Cimpian and Markman (2005) and Booth, Waxman, and Huang (2005). *Developmental Psychology, 42,* 1339–1343. <https://doi.org/10.1037/0012-1649.42.6.1339>

The British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: <http://www.natcorp.ox.ac.uk/>

Trueswell, J. C., & Papafragou, A. (2010). Perceiving and remembering events cross-linguistically: Evidence from dual-task paradigms. *Journal of Memory and Language, 63*, 64–82. <https://doi.org/10.1016/j.jml.2010.02.006>

Vanek, N., & Selinker, L. (2017). Covariation between temporal interlanguage features and nonverbal event categorisation. IRAL - *International Review of Applied Linguistics in Language Teaching, 55(3)*, 223–243. <https://doi.org/10.1515/iral-2017-0106>

Vendler, Z. (1957). Verbs and Times. *The Philosophical Review, 66(2)*, 143–160.

von Stutterheim, C., Andermann, M., Carroll, M., Flecken, M. & Schmiedtová, B. (2012). How grammaticized concepts shape event conceptualization in language production: Insights from linguistic analysis, eye tracking data, and memory performance. *Linguistics, 50(4),* 833–867. <https://doi.org/10.1515/ling-2012-0026>

Whorf, B. (1956). *Language, Thought, and Reality.* Cambridge, MA: MIT Press.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the United States of America, 104(19)*, 7780–7785. <https://doi.org/10.1073/pnas.0701644104>

Xiao, R., & McEnery, T. (2004). *Aspect in Mandarin Chinese: A corpus-based study.* Amsterdam: Benjamins. <https://doi.org/10.1075/slcs.73>

Yong, S. (1997). The grammatical functions of verb complements in Mandarin Chinese. *Linguistics, 35(1)*, 1-24. <https://doi.org/10.1515/ling.1997.35.1.1>