Author for correspondence:
Piera Filippi
e-mail: pie.filippi@gmail.com

[†]Present addresses: Brain and Language Research Institute, University Aix-Marseille, Marseille, France; Language and Cognition Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

THE ROYAL SOCIETY
PUBLISHING

Utterance-final position and pitch marking aid word learning in school-age children

Piera Filippi[1,†], Sabine Laaha[2] and W. Tecumseh Fitch[1]

[1]Department of Cognitive Biology, and [2]Department of Linguistics, University of Vienna, Vienna, Austria

PF, 0000-0002-2990-9344

We investigated the effects of word order and prosody on word learning in school-age children. Third graders viewed photographs belonging to one of three semantic categories while hearing four-word nonsense utterances containing a target word. In the *control* condition, all words had the same pitch and, across trials, the position of the target word was varied systematically within each utterance. The only cue to word–meaning mapping was the co-occurrence of target words and referents. This cue was present in all conditions. In the *Utterance-final* condition, the target word always occurred in utterance-final position, and at the same fundamental frequency as all the other words of the utterance. In the *Pitch peak* condition, the position of the target word was varied systematically within each utterance across trials, and produced with pitch contrasts typical of infant-directed speech (IDS). In the *Pitch peak + Utterance-final* condition, the target word always occurred in utterance-final position, and was marked with a pitch contrast typical of IDS. Word learning occurred in all conditions except the control condition. Moreover, learning performance was significantly higher than that observed with simple co-occurrence (*control* condition) only for the *Pitch peak + Utterance-final* condition. We conclude that, for school-age children, the combination of words' utterance-final alignment and pitch enhancement boosts word learning.

1. Introduction

A central issue in the study of language acquisition concerns the perceptual and memory constraints that human learners are subjected to as anchor points for word learning [1,2]. A number of experimental studies show that language learners are able to associate the sound of a word with its referent, at the moment the novel word is first encountered, or given repeated unambiguous

pairings in a single session [3,4]. However, word learning may go beyond the process of learning one label associated with one object [5]. In fact, language learners are typically exposed to many words and many potential referents, and need to use specific cues as to which sounds refer to which referents in the surrounding visual scene. Several studies show that infants rapidly learn multiple word–referent pairs by accruing statistical evidence across word–scene pairings [6,7]. Hence, a learner who is faced with multiple potential referents for a novel word on any single learning scene might store possible word–referent pairings across trials, i.e. evaluate the statistically regular co-occurrences between words and referents, and finally map individual words to their referents through this cross-trial evidence [8]. This kind of learning is shown to be sufficiently rapid and robust [6]. Indeed, humans are naturally predisposed to acquire words in their native language, a capacity that may involve the triple challenge of: (i) extracting (i.e. segmenting) and storing a word out of a continuous speech stream, (ii) inferring one or more possible referents within the visual scene and (iii) mapping the segmented word onto its referential/pragmatic meaning(s), and/or grammatical role(s) [9,10]. Linking the target word to its intended referent within the visual context is not a trivial task, as the number of potential referents in a complex scene is typically indefinite [11–13]. A final step in this process includes extending the acquired word over a potentially infinite set of appropriate referents within the same semantic category, for instance using the word 'dog' for dogs that the given learner has never seen before [5,14], and to an open-ended set of novel utterances [15].

A variety of cues are known to assist language learners in each of these processes involved in word learning [1,16]. Studies suggest that word frequency [17], phonotactic regularities [18–22], tactile cues [23] and prosodic modulation [24,25] orient word segmentation in infants. Importantly, carers across multiple cultures use a special speech register (infant-directed speech or 'parentese') characterized by elevated fundamental frequency (pitch) and intonational contours [26,27], hyperarticulated vowels ([28,29], but see recent contrasting evidence [30–32]), and high emotional content [33]. These acoustic adjustments in speech are shown to engage attention in the communication act, facilitating word segmentation for both infants [34,35] and adults [36]. In addition, a critical cue in word segmentation is the position of an item, which can also affect word acquisition. Research suggests that the general property of primacy and recency in memory abilities may focus attention on words at the edges of an utterance in infants [37–39] and adults [36,40,41]. Seidl & Johnson [38] found that infants segment words from the edges of utterances more readily than from the middle of utterances [1,39,40]. In other words, memory biases may constrain the processing of auditory input, allocating the learners' cognitive resources to the edges of the auditory utterances [42]. Although cognitive bias to the edges of utterances might be related to prosodic modifications such as initial strengthening and final lengthening of words at utterance boundaries [39], very little attention has been paid to the effect of other prosodic cues on learning words at utterance boundaries. Crucially, a number of studies suggest that statistical regularities in words guide categorization of visual objects [43–46]. Finally, as to the word–referent mapping process, research suggests that statistical patterns within words [47,48] or phrases [44,49], and the speaker's eye gaze [50] guide word–object associations in infants. In addition, much research suggests that prosody is particularly important in word–referent mapping, as it guides the infants' attention to the relevant speech component and the corresponding visual input [51], affecting their ability to acquire the intended referent [5,37,52–55]. However, in contrast to research on word segmentation and word–object fast mapping, little research has addressed perceptual cues affecting word–meaning mapping in a cross-situational task, and the extension of acquired words over novel appropriate referents [5,56,57]. A recent study by Filippi et al. [58] suggests that word learning is aided in adults when target words, which are consistently paired with referents across multiple visual contexts, are marked with a pitch contrast typical of infant-directed speech (IDS). Furthermore, this study suggested that pitch enhancement has a stronger effect than other possible visual and acoustic perceptual spotlights in aiding word learning in adults.

In this work, we disentangled, for the first time, the relative effect of two critical cues on word learning, prosody and position of the target words in a sentence, within a cross-situational statistical word learning task. Specifically, we tested the prediction that pitch enhancement typical of IDS and word position at the final edge of utterances significantly support word learning across multiple visual contexts, both alone and when combined.

Surprisingly, although much research has investigated the effect of perceptual salience markers in aiding word learning in infants and adults, no research we know of has addressed this effect in school-aged children. Examining subjects at this age of rapid and steady vocabulary growth is particularly relevant to enhance our understanding of the developmental curve of language learning from infancy through child- and adulthood. Critically, in this study, we tested third graders (8- to 9-year-olds) in a

cross-situational word learning task. Our work provides empirical evidence suggesting that positioning a target word in utterance-final position and/or marking it with a pitch emphasis typical of IDS enables word learning in third graders. We found that the combination of these two cues significantly boosts word learning.

2. Material and methods

2.1. Participants

A total of 56 children (26 female, mean age = 8;11, range: 8–10; s.d. = 6.7 months) took part in the study. They were all monolingual native speakers of German and were recruited from an elementary school in a middle-to-high socio-economic neighbourhood in Vienna (Austria).

Exclusion criteria included bilingualism, known developmental learning disorders, and visual or auditory impairment. Participants were given colouring pens in exchange for their participation.

2.2. Material

In this study, we adopted the EIM (target sound string Extraction, referential category Inference and word–meaning Mapping) task developed in Filippi *et al.* [59]. This paradigm uses complex naturalistic images of target objects, providing a realistic visual parsing challenge.

The stimuli consisted of photographic images, presented on an LCD touchscreen monitor, paired with artificial language utterances presented over headphones. These spoken utterances contain a target word, which is associated with the intended image category (hereafter 'semantic category') of the photograph (figure 1). The co-occurrence between target words and their intended referent was the only consistent statistical cue that could help participants learn the target word–referent mapping. Participants had to identify the target words across the utterances, infer the intended semantic category from the photographs, and link these two together into a word–meaning pair which allows them to subsequently extend the acquired word to novel utterances and to new visual contexts (novel images). Custom software (EXPERIMENTER v. 3.5) written in Python v. 2.6 was used to present the stimuli and collect touchscreen responses.

2.2.1. Images

Thirty-six unique full-colour images of real-life scenes were selected each depicting one of three intended visual semantic categories: 'dog', 'tree' and 'ball'. The images were downloaded from 'Creative Commons' websites and scaled to 300 × 300 pixels. Care was taken that no obvious emotional or written content was depicted in these pictures.

2.2.2. Sounds

We invented an artificial language made of 36 four-word utterances (hereafter 'utterances'). Each word was a CV (consonant + vowel) monosyllabic unit. Thus, each utterance was a string of four nonsense monosyllabic 'words', including the target word (table 1; sound samples in electronic supplementary material). The CV syllables came from a pool of five vowels (a, e, i, o, u) and 12 consonants, namely four stops (p, t, k, g), four fricatives (f, v, s, sh) and four sonorants (m, n, l, r). Care was taken that no words or parts of utterances of our artificial language coincided with real words of the language spoken by the participants (German). As shown in table 1, in each utterance, across all conditions, each monosyllabic word occurred only once.

The artificial language was subdivided into three different sets of 12 utterances, each of which contained the target word that referred to one of the three semantic categories depicted in the images (as in Filippi [58]). Specifically, each set of utterances shared one distinctive word, the target word, which consistently occurred in association with the corresponding visual semantic category. Hence, there were three semantic sets of utterances, i.e. sets of utterances corresponding to a specific visual semantic category (dog, tree or ball). All non-target words were systematically shared across utterances of the three semantic categories, and therefore had no consistent referential link to the visual stimuli (table 1). Within each learning condition, all non-target words were identical and systematically occurred in the same position across semantic categories. Consider, for instance, the utterance 'loFUsheti'. The 'lo' and 'sheti' parts of the utterance occurred also within the utterances 'loGAsheti' and 'loMIsheti'. Hence, only the monosyllabic words shared within each semantic set of utterances constituted target
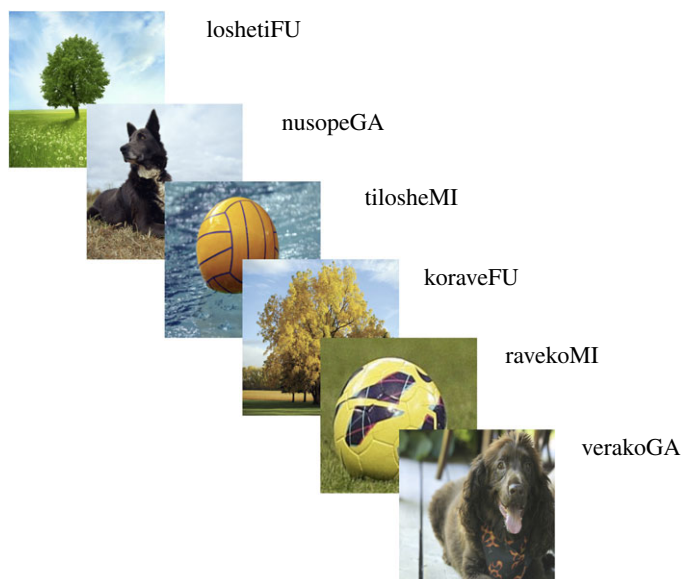
Figure 1. Example stimuli presentation series in the *Pitch peak* + *Utterance-final* condition. In each experimental condition, participants were exposed to 36 successive stimuli, consisting of images paired with an auditory utterance of four monosyllabic words. Each image category—dog, tree, ball—was linked only to a specific word (target word) randomly assigned to that referential category (different for different subjects). In this example, /fu/ always co-occurs with the category 'tree', /ga/ with 'dog' and /mi/ with 'ball' (capitalized in the figure).

words mappable to a specific visual semantic category. Across learning conditions, within each semantic category, forward transitional probabilities between monosyllabic words, i.e. the probability of one word given the occurrence of the preceding word within each utterance, were 0, 0.25 or 0.50.

We wanted to avoid that children who were tested would reveal the correct word–meaning associations to children who still had to go through the experiment. To this goal, we adopted three different sets of target words and utterances (all built according to the description above), which were randomly assigned across participants (table 1). The three sets of target words were the following: (i) fu/ga/mi; (ii) na/tu/vi and (iii) pi/ru/sa.

In order to avoid co-articulation between adjacent words, and following Filippi *et al*. [59], each word was recorded individually. Acoustic parameters of each word were then modified using PRAAT [60]. In particular, the words' pitch and duration were modulated using the pitch-synchronous overlap add (PSOLA) algorithm [61]. PSOLA is a method based on decomposition of a signal into a series of elementary waveforms in such a way that each waveform represents one of the successive pitch periods of the signal and the sum (overlap add) of them reconstitutes the signal. The basic algorithm for the PSOLA technique consists of three steps. First, the speech waveform is divided into smaller, short-term analysis segments. Second, a mathematical function is applied to the signal, centring it on the successive instants t_m, called 'pitchmarks'. These marks are set at a pitch-synchronous rate on the voiced parts of the signal and at a constant rate on the unvoiced parts. Finally, these segments are modified, by either repeating or leaving out speech segments, depending on whether the target fundamental frequency of the signal is higher or lower than the fundamental frequency of the source signal. The remaining segments are recombined into a synthesized signal through overlapping and adding. The synthesized signal has the same spectral envelop as the source signal, but a different fundamental frequency. To change the duration of the signal, the speech segments may be repeated multiple times, to increase the duration—or eliminated, to decrease the duration. The segments are then combined into a synthesized signal using the overlap add technique.

In the Control and Utterance-final conditions (see 'Learning conditions' section), pitch, loudness and duration were normalized for all words within each utterance. The monosyllabic words were concatenated without pauses to form four-word utterances. Here, the target words' pitch was normalized to have the same F_0 as the three other words ($M = 210.2$ Hz; s.d. $= 0.22$ Hz). In the Pitch peak and Pitch peak + Utterance-final conditions, loudness and duration were normalized for all words within each utterance. In these two learning conditions, in order to highlight the perceptual salience of the target word, its pitch value was elevated ($M = 314.23$ Hz; s.d. $= 5.97$ Hz; Max $= 419.45$ Hz; s.d. $= 0.55$ Hz). We

Table 1. Utterances included in our artificial language, subdivided by semantic category (dog, tree, ball) and learning condition. Target words are capitalized. Each target word was randomly assigned to a different semantic category for each subject. We adopted three different sets of target words and utterances, which were randomly assigned across participants.

	target words: 'FU', 'GA', 'MI'							
	training				test			
conditions: *Utterance-final, Pitch peak + Utterance-final condition*								
semantic category A	pesonuFU	nusopeFU	sonupeFU	penusoFU	kovetiFU	tivekoFU	vetikoFU	kotiveFU
	tisheloFU	loshetiFU	shelotiFU	tilosheFU	nusheloFU	loshenuFU	shelonuFU	nulosheFU
	koveraFU	ravekoFU	verakoFU	koraveFU	rasopeFU	pesoraFU	soperaFU	rapesoFU
semantic category B	pesonuGA	nusopeGA	sonupeGA	penusoGA	kovetiGA	tivekoGA	vetikoGA	kotiveGA
	tisheloGA	loshetiGA	shelotiGA	tilosheGA	nusheloGA	loshenuGA	shelonuGA	nulosheGA
	koveraGA	ravekoGA	verakoGA	koraveGA	rasopeGA	pesoraGA	soperaGA	rapesoGA
semantic category C	pesonuMI	nusopeMI	sonupeMI	penusoMI	kovetiMI	tivekoMI	vetikoMI	kotiveMI
	tisheloMI	loshetiMI	shelotiMI	tilosheMI	nusheloMI	loshenuMI	shelonuMI	nulosheMI
	koveraMI	ravekoMI	verakoMI	koraveMI	rasopeMI	pesoraMI	soperaMI	rapesoMI
conditions: control condition, *Pitch peak condition*								
semantic category A	FUpesonu	nuFUsope	sonuFUpe	penusoFU	FUkoveti	tiFUveko	vetiFUko	kotiveFU
	FUtishelo	loFUsheti	sheloFUti	tilosheFU	FUnushelo	loFUshenu	sheloFUnu	nulosheFU
	FUkovera	raFUveko	veraFUko	koraveFU	FUrasope	peFUsora	sopeFUra	rapesoFU
semantic category B	GApesonu	nuGAsope	sonuGApe	penusoGA	GAkoveti	tiGAveko	vetiGAko	kotiveGA
	GAtishelo	loGAsheti	sheloGAti	tilosheGA	GAnushelo	loGAshenu	sheloGAnu	nulosheGA
	GAkovera	raGAveko	veraGAko	koraveGA	GArasope	peGAsora	sopeGAra	rapesoGA
semantic category C	MIpesonu	nuMIsope	sonuMIpe	penusoMI	MIkoveti	tiMIveko	vetiMIko	kotiveMI
	MItishelo	loMIsheti	sheloMIti	tilosheMI	MInushelo	loMIshenu	sheloMInu	nulosheMI
	MIkovera	raMIveko	veraMIko	koraveMI	MIrasope	peMIsora	sopeMIra	rapesoMI
	target words: 'NA', 'TU', 'VI'							
	training				test			
conditions: *Utterance-final, Pitch peak + Utterance-final condition*								
semantic category A	pesomuNA	musopeNA	somupeNA	pemusoNA	kofegiNA	gifekoNA	fegikoNA	gikofeNA
	gisheloNA	loshegiNA	shelogiNA	gilosheNA	shemuloNA	loshemuNA	shelomuNA	mulosheNA
	koferaNA	rafekoNA	ferakoNA	korafeNA	rasopeNA	pesoraNA	soperaNA	rapesoNA
semantic category B	pesomuTU	musopeTU	somupeTU	pemusoTU	kofegiTU	gifekoTU	fegikoTU	gikofeTU
	gisheloTU	loshegiTU	shelogiTU	gilosheTU	shemuloTU	loshemuTU	shelomuTU	mulosheTU
	koferaTU	rafekoTU	ferakoTU	korafeTU	rasopeTU	pesoraTU	soperaTU	rapesoTU
semantic category C	pesomuVI	musopeVI	somupeVI	pemusoVI	kofegiVI	gifekoVI	fegikoVI	gikofeVI
	gisheloVI	loshegiVI	shelogiVI	gilosheVI	shemuloVI	loshemuVI	shelomuVI	mulosheVI
	koferaVI	rafekoVI	ferakoVI	korafeVI	rasopeVI	pesoraVI	soperaVI	rapesoVI
conditions: control condition, *Pitch peak condition*								
semantic category A	NApesomu	muNAsope	somuNApe	pemusoNA	NAkofegi	giNAfeko	fegiNAko	gikofeNA
	NAgishelo	loNAshegi	sheloNAgi	gilosheNA	NAshemulo	loNAshemu	sheloNAmu	mulosheNA
	NAkofera	raNAfeko	feraNAko	korafeNA	NArasope	peNAsora	sopeNAra	rapesoNA
semantic category B	TUpesomu	muTUsope	somuTUpe	pemusoTU	TUkofegi	giTUfeko	fegiTUko	gikofeTU
	TUgishelo	loTUshegi	sheloTUgi	gilosheTU	TUshemulo	loTUshemu	sheloTUmu	mulosheTU
	TUkofera	raTUfeko	feraTUko	korafeTU	TUrasope	peTUsora	sopeTUra	rapesoTU

(*Continued.*)

Table 1. (*Continued.*)

	target words: 'NA', 'TU', 'VI'							
	training				test			
semantic category C	VIpesomu	muVIsope	somuVIpe	pemusoVI	VIkofegi	giVIfeko	fegiVIko	gikofeVI
	VIgishelo	loVIshegi	sheloVIgi	gilosheVI	VIshemulo	loVIshemu	sheloVImu	mulosheVI
	VIkofera	raVIfeko	feraVIko	korafeVI	VIrasope	peVIsora	sopeVIra	rapesoVI
	Target words: 'PI', 'RU', 'SA'							
	training				test			
conditions: *Utterance-final, Pitch peak + Utterance-final condition*								
semantic category A	gefonuPI	nufogePI	fonugePI	genufoPI	kovetiPI	tivekoPI	vetikoPI	kotivePI
	tisheloPI	loshetiPI	shelotiPI	tiloshePI	nusheloPI	loshenuPI	shelonuPI	nuloshePI
	kovemaPI	mavekoPI	vemakoPI	makovePI	mafogePI	gefomaPI	fogemaPI	magefoPI
semantic category B	gefonuRU	nufogeRU	fonugeRU	genufoRU	kovetiRU	tivekoRU	vetikoRU	kotiveRU
	tisheloRU	loshetiRU	shelotiRU	tilosheRU	nusheloRU	loshenuRU	shelonuRU	nulosheRU
	kovemaRU	mavekoRU	vemakoRU	makoveRU	mafogeRU	gefomaRU	fogemaRU	magefoRU
semantic category C	gefonuSA	nufogeSA	fonugeSA	genufoSA	kovetiSA	tivekoSA	vetikoSA	kotiveSA
	tisheloSA	loshetiSA	shelotiSA	tilosheSA	nusheloSA	loshenuSA	shelonuSA	nulosheSA
	kovemaSA	mavekoSA	vemakoSA	makoveSA	mafogeSA	gefomaSA	fogemaSA	magefoSA
conditions: control condition, *Pitch peak condition*								
semantic category A	PIgefonu	nuPIfoge	fonuPIge	genufoPI	PIkoveti	tiPIveko	vetiPIko	kotivePI
	PItishelo	loPIsheti	sheloPIti	tiloshePI	PInushelo	loPIshenu	sheloPInu	nuloshePI
	PIkovema	maPIveko	vemaPIko	makovePI	PImafoge	gePIfoma	fogePIma	magefoPI
semantic category B	RUgefonu	nuRUfoge	fonuRUge	genufoRU	RUkoveti	tiRUveko	vetiRUko	kotiveRU
	RUtishelo	loRUsheti	sheloRUti	tilosheRU	RUnushelo	loRUshenu	sheloRUnu	nulosheRU
	RUkovema	maRUveko	vemaRUko	makoveRU	RUmafoge	geRUfoma	fogeRUma	magefoRU
semantic category C	SAgefonu	nuSAfoge	fonuSAge	genufoSA	SAkoveti	tiSAveko	vetiSAko	kotiveSA
	SAtishelo	loSAsheti	sheloSAti	tilosheSA	SAnushelo	loSAshenu	sheloSAnu	nulosheSA
	SAkovema	maSAveko	vemaSAko	makoveSA	SAmafoge	geSAfoma	fogeSAma	magefoSA

adopted a pitch deviation of one octave, a pitch excursion typical of IDS across multiple languages [26]. Adopting a pitch exaggeration typically employed in IDS allows us to analyse the effect of highly perceptible pitch enhancement in the process of word learning. Mean intensity of target and non-target monosyllabic words was 69.97 dB (s.d. = 0.08) relative to peak amplitude. The duration of each word was normalized ($M = 404$ ms; s.d. = 3 ms).

2.3. Learning conditions

We adopted a between-subjects design; 14 participants were included in each of the following four learning conditions.

(1) *Control condition*: the position of the target word was varied systematically across each utterance, appearing in each of the four 'slots' with equal frequency. Here, the only cue that could support successful word learning would be the consistent co-occurrence of the target word with the corresponding semantic category (hereafter 'statistical cue'). This level of information is present in all experimental conditions.

(2) *Utterance-final condition*: identical to the control condition, with the exception that the target word always occurred in the last slot in each utterance.

(3) *Pitch peak condition*: the pitch contrast manipulation described above was used in addition to the statistical cue. The position of the target word was varied systematically across each utterance, appearing in each of the four 'slots' with equal frequency.

(4) *Pitch peak + Utterance-final condition*: identical to the pitch peak condition, with the exception that the target word always occurred in the last slot in each utterance.

2.4. Training and testing procedure

Children were tested individually in a quiet room at their school, in a single session of approximately 15 min. Experimenters sat in the same room, at approximately 2 m from the participants.

An explicit learning paradigm was used. Participants were instructed through a cartoon-video created on *goanimate.com*. Here, a cartoon alien explained that they could participate in an 'Alien Language Learning Game' in which they would see a series of pictures and hear the sounds that the imaginary alien would use to describe those pictures. Participants were instructed that the experiment consisted of a training and a test phase. They were asked to do their best to understand as much as they could of this 'alien' language in the training, and told that their mastery of the language would be assessed in a test phase following the training. They were instructed that prior to the experiment start, they would run a practice phase in order to familiarize themselves with the 'game' procedure. As the actual experiment, the practice phase included training and a test phase. The practice-training consisted of six trials in which German utterances adjusted to a monotone pitch were played in association with a target image. The practice-test consisted of six trials, in which three images were shown and a monotone German utterance played. Here, children were instructed to touch the image that the sentence referred to. Two children, who did not answer correctly to more than three test trials in the practice phase, thus showing misunderstanding of the experimental procedure or lack of motivation, were excluded from the data analysis.

In both the training and the test phases of the actual experiment, the artificial language was manipulated as described in the 'Learning conditions' section above. Each target word was randomly assigned to a different semantic category for each subject. The training session consisted of 36 exposure trials and lasted approximately two minutes. Each utterance, and each image, was presented only once. The auditory unit–image pairs were presented in a random order across participants. For each slide, the onset of the utterance was aligned with the onset of image presentation. Consequently, only in the control and pitch peak conditions, where the target word's position was varied systematically across each utterance, onset of the target word coincided with onset of image presentation in 9 of 36 trials. In all training trials, the image remained on screen for a further 1500 ms after the end of the auditory unit's presentation.

After the training session, participants received a multiple-choice test. In the test phase, participants were presented with a novel four-word utterance, containing one of the three target words. Here, we used 12 novel utterances per semantic category, for a total of 36 novel utterances. In each test trial, one utterance was played, and three images were simultaneously shown on screen (again, one image per semantic category). Each utterance was associated once with a set of three probe images presented simultaneously, yielding 36 test trials. The onset of images presentation coincided with the onset of the auditory utterance. Participants were asked to indicate which image matched the auditory unit by touching one of the three images on the screen. The images remained on screen for 4000 ms after the end of the auditory unit's presentation. Thus, participants could thus make their choice from the end of the auditory stimulus playback to up to 4 s after the sound ended. No feedback was provided. An interval of 1000 ms followed the subject's response on each trial prior to the onset of the next trial. The order of presentation of the utterance-image trials, as well as the left-to-right arrangement of the three images on the monitor was randomized for each subject. Test utterances were the same for all participants within each learning condition. Presenting novel images probes the participants' ability to apply the acquired word to a potentially infinite set of new instances of its intended reference. The novel utterances examined their ability to identify the acquired word within an open-ended set of new utterances.

3. Results

Statistical analyses were performed using SPSS for Mac OS X v. 19.

We excluded timeouts from the analyses because they could not be analysed as either correct or incorrect responses. A binomial test revealed that participants performed significantly better than chance (33.3%) in all conditions (Utterance-final and Pitch peak: $p < 0.01$, one-tailed; Pitch peak + Utterance-final condition: $p < 0.001$, one-tailed) except the control condition ($p = 0.344$, one-tailed) (figure 2). A binary logistic regression model was built within the generalized linear model framework, to compare responses
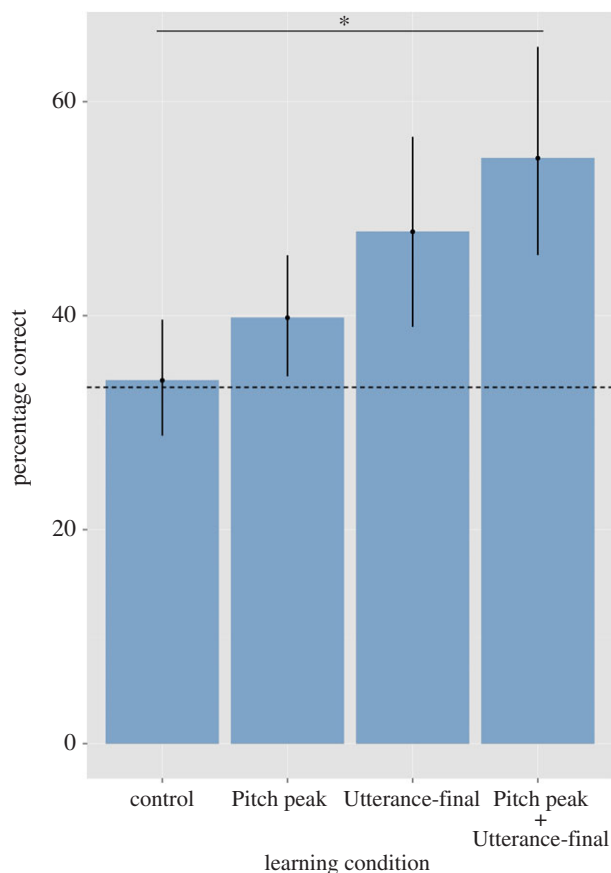
Figure 2. Percentage of correct responses in each experimental condition. Error bars represent 95% confidence intervals. Chance performance level is set at 33% (dashed line). All conditions except the control condition were significantly better than chance. The horizontal line (*) indicates the significant pairwise comparison between the control condition and the Pitch peak + Utterance-final condition ($p = 0.037$).

across learning conditions. Data across all subjects were modelled using a binomial distribution and a logit link function. Semantic category was entered as a within-subject predictor variable and learning condition as a between-group predictor variable. The dependent variable was the proportion of correct choices in participants' responses (where chance = 33.3%) over the number of test trials for which a response was entered. In other words, the number of trials was calculated by subtracting the number of total test trials included in our design, i.e. 12 per semantic category, minus the number of timeouts. The model revealed a significant main effect of learning condition (Wald $\chi_3^2 = 8.823$, $p = 0.032$), no significant effect of semantic category (Wald $\chi_2^2 = 4.439$, $p = 0.109$) and no significant interactions between semantic category and learning condition (Wald $\chi_6^2 = 6.940$, $p = 0.326$). Pairwise comparisons between the control condition and all the other learning conditions, using Bonferroni's correction, revealed a significant difference only between the control condition and the Pitch peak + Utterance-final condition (Wald $\chi_1^2 = 6.256$, $p = 0.037$; Cohen's $d = 0.75$; $r = 0.35$). Differences in learning performance did not reach significance between the control and the Utterance-final conditions (Wald $\chi_1^2 = 3.460$, $p = 0.189$; Cohen's $d = 0.56$; $r = 0.27$), and between the control and the Pitch peak conditions (Wald $\chi_1^2 = 1.615$, $p = 0.611$; Cohen's $d = 0.25$; $r = 0.12$) (figure 2).

4. Discussion

Word learning includes multiple processes: word segmentation and storage, reference induction from the visual context(s) and word–referent mapping. A number of cues and perceptual salience markers as, for instance, word position within the utterance and prosodic modulation of the signal, aid the language learners in this task. Prior studies have addressed the effect of both these cues in infants [26,35,37–39,42,52] and of pitch enhancement in adults [37,59]. However, to our knowledge little attention has

been given to the role of perceptual and attentional highlight markers typical of IDS in children. This study was designed to determine the effect of words' pitch contrasts of IDS and Utterance-final position in facilitating word learning in school-age children, who are still very actively acquiring new vocabulary.

We found that 8- to 9-year-old children can acquire three new word–meaning mappings in approximately 2 min (36 exposure trials) in all conditions except the control condition (co-occurrence alone), which did not provide any perceptual or positional cues for learning. This suggests that the co-occurrence of target word and its intended referent with only one of the two salience markers we examined in this study is sufficient for word learning in school-age children. However, among the three successful conditions, only participants in the Pitch peak + Utterance-final condition reached a level of learning performance that significantly differed from the learning performance in the control condition. Hence, our data suggest that performance in this EIM task is significantly boosted when the two cues (words' pitch enhancement and Utterance-final position) are combined.

Being limited to the examination of pitch exaggeration and Utterance-final position, this study leaves open the effect of other words' prosodic marking, or of Utterance-initial position. We may speculate that further cues aiding memory, for instance placing the target word in Utterance-initial position—if combined with cross-situational co-occurrence of target word and visual semantic category and with pitch marking of the target word—would ease word learning in school-age children.

Our findings are in line with other studies [36,38,39,49], suggesting that edge alignment facilitates word learning in preverbal infants. Our work extends these studies, suggesting that recency effects alone may not be sufficient to ease word segmentation, and that prosodic modifications of the words in Utterance-final position are necessary to facilitate word learning. Our results are particularly interesting if linked to findings described in Filippi *et al.* [59] on adults, where an identical experimental paradigm was adopted. Filippi *et al.* [59] suggested that the word–reference co-occurrence cue alone is sufficient for word learning in adults. In addition, the authors found that word learning is boosted only when pitch enhancement marks the target word, also when the position of target words varied across utterances. A possible explanation for the discrepancy between the present research and Filippi *et al.* [59] may be due to differences in adults' and children's sensitivity to salient features in the spoken utterances or to different cognitive resources for word learning in children as compared to adults [55]. It can be argued that, in contrast to word learning in adults, word learning in school-age children is significantly boosted when multiple cues and/or salient perceptual modifications mark target words. In fact, our results complement previous findings on word learning, showing that processing multiple sources of information (e.g. unfamiliar, but statistically regular phonetic variations and word–objects associations) at the same time supports cross-modal learning [46,62,63]. Our data suggest that for school-age children, the combination of *specific* cues aiding memory and enhancing word perception—namely cross-situational statistical regularities between acoustic signal and visual image, pitch enhancement and utterance-final position—facilitate word learning. More work will be needed to explore the combination of other types of cues enhancing the target word (e.g. word lengthening, or vowel hyperarticulation) with word position in utterances within a cross-situational word learning task. Crucially, further work disentangling the role of statistical regularities within the utterances and the relative contribution of other learning cues is suggested.

In addition, future research should aim at improving the experimental design making the task easier for children, possibly testing also younger participants on the same task. This may be achieved, for instance, by implementing the experiment into a videogame paradigm—thus making the task more involving for children [64,65], by including more acoustic variation, or by elongating the training phase. This work specifically focuses on word learning, which we define as including generalizations of target words used in previously unheard utterances, and applied to novel images. Further work is required to establish whether children would succeed in recognizing the target words in a simpler test task, where, for instance the same images used in the training are paired with novel utterances. This line of research will contribute to develop a full picture of factors aiding all the different processes and mechanisms involved in language learning across different stages of the language development curve. Furthermore, future research comparing children's and adults' performance in language learning tasks should examine whether children are more sensitive than adults to word position, perhaps using an artificial language where word position and/or prosodic marking is linked to syntactic properties in the signal.

Our results extend research concerning the beneficial effect of perceptual marking for tasks that engage memory and attentional resources, as was the case for our word learning task [1,5,44,45,51,54]. Moreover, the present study contributes additional evidence suggesting that 8- to 9-year-old language learners exploit pitch enhancement and word Utterance-final position as cues that co-occur with the

intended target words and visual categories across learning contexts. These findings align with previous research on natural languages, underlining the importance of positional regularities for language acquisition and processing [1].

This research may be relevant for educational programmes aiming to facilitate the process of second language(s) learning, or to investigate the specific uses of voice modulation as a teaching strategy in school-age children. Further research addressing the effect of word markers in enhancing children's language learning will improve our understanding of the processes underlying language development and acquisition.

References

1. Endress AD, Nespor M, Mehler J. 2009 Perceptual and memory constraints on language acquisition. *Trends Cogn. Sci.* **13**, 348–353. (doi:10.1016/j.tics.2009.05.005)

2. Endress AD, Mehler J. 2010 Perceptual constraints in phonotactic learning. *J. Exp. Psychol. Hum. Percept. Perform.* **36**, 235–250. (doi:10.1037/a0017164)

3. Woodward AL, Markman EM, Fitzsimmons CM. 1994 Rapid word learning in 13- and 18-month-olds. *Dev. Psychol.* **30**, 553–566. (doi:10.1037/0012-1649.30.4.553)

4. Schafer G, Plunkett K. 1998 Rapid word learning by fifteen-month-olds under tightly controlled conditions. *Child Dev.* **69**, 309–320. (doi:10.1111/j.1467-8624.1998.tb06190.x)

5. Waxman SR, Gelman SA. 2009 Early word-learning entails reference, not merely associations. *Trends Cogn. Sci.* **13**, 258–263. (doi:10.1016/j.tics.2009.03.006)

6. Smith L, Yu C. 2008 Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition* **106**, 1558–1568. (doi:10.1016/j.cognition.2007.06.010)

7. Yu C, Smith LB. 2007 Rapid word learning under uncertainty via cross-situational statistics: research article. *Psychol. Sci.* **18**, 414–420. (doi:10.1111/j.1467-9280.2007.01915.x)

8. Trueswell JC, Medina TN, Hafri A, Gleitman LR. 2013 Propose but verify: fast mapping meets cross-situational word learning. *Cogn. Pyscol.* **66**, 126–156. (doi:10.1016/j.cogpsych.2012.10.001)

9. Baldwin D. 1993 Early referential understanding: infants' ability to recognize referential acts for what they are. *Dev. Psychol.* **29**, 832–843. (doi:10.1037/0012-1649.29.5.832)

10. Bloom P. 2000 *How children learn the meanings of words*. Cambridge, MA: MIT Press.

11. Gleitman L. 1990 The structural sources of verb meanings. *Lang. Acquis.* **1**, 3–55. (doi:10.1207/s15327817la0101_2)

12. Medina TN, Snedeker J, Trueswell JC, Gleitman LR. 2011 How words can and cannot be learned by observation. *Proc. Natl Acad. Sci. USA* **108**, 9014–9019. (doi:10.1073/pnas.1105040108)

13. Quine W. 1960 *Word and object*. Cambridge, MA: MIT Press.

14. Brown R. 1958 *Words and things*. Glencoe, IL: Free Press.

15. Chomsky N. 2000 *New horizons in the study of language and mind*. Cambridge, UK: Cambridge University Press.

16. Gómez R, Gerken L. 2000 Infant artificial language learning and language acquisition. *Trends Cogn. Sci.* **4**, 178–186. (doi:10.1016/S1364-6613(00)01467-4)

17. Frost RLA, Christiansen MH. 2016 Using statistics to learn words and grammatical categories: how high frequency words assist language acquisition. In *Proc. 38th Annu. Conf. Cognitive Science Society* (eds A Papafragou, D Grodner, D Mirman, JC Trueswell), pp. 81–86. Austin, TX: Cognitive Science Society.

18. Saffran JR, Aslin RN, Newport EL. 1996 Statistical learning by 8-month-old infants. *Science* **274**, 1926–1928. (doi:10.1126/science.274.5294.1926)

19. Pelucchi B, Hay JF, Saffran JR. 2009 Statistical learning in a natural language by 8-month-old infants. *Child Dev.* **80**, 674–685. (doi:10.1111/j.1467-8624.2009.01290.x)

20. Mattys SL, Jusczyk PW, Luce PA, Morgan JL. 1999 Phonotactic and prosodic effects on word segmentation in infants. *Cogn. Psychol.* **38**, 465–494. (doi:10.1006/cogp.1999.0721)

21. Wagner M, Watson DG. 2010 Experimental and theoretical advances in prosody: a review. *Lang. Cogn. Process.* **25**, 905–945. (doi:10.1080/01690961003589492)

22. Jusczyk PW, Friederici AD, Wessels JM, Svenkerud VY, Jusczyk AM. 1993 Infants' sensitivity to the sound patterns of native language words. *J. Mem. Lang.* **32**, 402–420. (doi:10.1006/jmla.1993.1022)

23. Seidl A, Tincoff R, Baker C, Cristia A. 2015 Why the body comes first: effects of experimenter touch on infants' word finding. *Dev. Sci.* **18**, 155–164. (doi:10.1111/desc.12182)

24. Johnson EK, Jusczyk PW. 2001 Word segmentation by 8-month-olds: when speech cues count more than statistics. *J. Mem. Lang.* **44**, 548–567. (doi:10.1006/jmla.2000.2755)

25. Johnson EK, Seidl AH. 2009 At 11 months, prosody still outranks statistics. *Dev. Sci.* **12**, 131–141. (doi:10.1111/j.1467-7687.2008.00740.x)

26. Fernald A. 1992 Human maternal vocalizations to infants as biologically relevant signals: an evolutionary perspective. In *The adapted mind: evolutionary psychology and the generation of culture* (eds JH Barkow, L Cosmides, J Tooby), pp. 391–428. Oxford, UK: Oxford University Press.

27. Fernald A, Kuhl P. 1987 Acoustic determinants of infant preference for motherese speech. *Infant Behav. Dev.* **10**, 279–293. (doi:10.1016/0163-6383(87)90017-8)

28. Burnham D, Kitamura C, Vollmer-Conna U. 2002 What's new, pussycat? On talking to babies and animals. *Science* **296**, 1435. (doi:10.1126/science.1069587)

29. Kuhl PK, Andruski JE, Chistovich IA, Chistovich LA, Kozhevnikova EV, Ryskina VL, Stolyarova EI, Sundberg U, Lacerda F. 1997 Cross-language analysis of phonetic units in language addressed to infants. *Science* **277**, 684–686. (doi:10.1126/science.277.5326.684)

30. Martin A, Schatz T, Versteegh M, Miyazawa K, Mazuka R, Dupoux E, Cristia A. 2015 Mothers speak less clearly to infants than to adults: a comprehensive test of the hyperarticulation hypothesis. *Psychol. Sci.* **26**, 341–347. (doi:10.1177/0956797614562453)

31. McMurray B, Kovack-Lesh KA, Goodwin D, McEchron W. 2013 Infant directed speech and the development of speech perception: enhancing development or an unintended consequence? *Cognition* **129**, 362–378. (doi:10.1016/j.cognition.2013.07.015)

32. Cristia A, Seidl A. 2014 The hyperarticulation hypothesis of infant-directed speech. *J. Child Lang.* **41**, 913–934. (doi:10.1017/S0305000912000669)

33. Trainor LJ, Austin CM, Desjardins RN. 2000 Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychol. Sci.* **11**, 188–195. (doi:10.1111/1467-9280.00240)

34. Thiessen E, Hill E, Saffran J. 2005 Infant-directed speech facilitates word segmentation. *Infancy* **7**, 53–71. (doi:10.1207/s15327078in0701_5)

35. Floccia C, Keren-Portnoy T, DePaolis R, Duffy H, Delle Luche C, Durrant S, White L, Goslin J, Vihman M. 2016 British English infants segment words only with exaggerated infant-directed speech stimuli. *Cognition* **148**, 1–9. (doi:10.1016/j.cognition.2015.12.004)

36. Golinkoff RM, Alioto A. 1995 Infant-directed speech facilitates lexical learning in adults hearing Chinese: implications for language acquisition. *J. Child Lang.* **22**, 703–726. (doi:10.1017/S030500090001011)

37. Fernald A, Mazzie C. 1991 Prosody and focus in speech to infants and adults. *Dev. Psychol.* **27**, 209–221. (doi:10.1037/0012-1649.27.2.209)

38. Seidl A, Johnson EK. 2006 Infant word segmentation revisited: edge alignment facilitates target extraction. *Dev. Sci.* **9**, 565–573. (doi:10.1111/j.1467-7687.2006.00534.x)

39. Seidl A, Johnson EK. 2008 Boundary alignment enables 11-month-olds to segment vowel initial words from speech. *J. Child Lang.* **35**, 1–24. (doi:10.1017/S0305000907008215)

40. Baddeley AD, Hitch G. 1993 The recency effect: implicit learning with explicit retrieval? *Mem. Cognit.* **21**, 146–155. (doi:10.3758/BF03202726)

41. Baddeley A. 1996 The fractionation of working memory. *Proc. Natl Acad. Sci. USA* **93**, 13 468–13 472. (doi:10.1073/pnas.93.24.13468)

42. Slobin DI. 1973 Cognitive prerequisites for the development of grammar. In *Studies of child language development* (eds CA Ferguson, DI Slobin), pp. 175–208. New York, NY: Holt, Rinehart and Winston.

43. Sloutsky VM, Robinson C. 2008 The role of words and sounds in infants' visual processing: from overshadowing to attentional tuning. *Cogn. Sci. A Multidiscip. J.* **32**, 342–365. (doi:10.1080/03640210701863495)

44. Napolitano AC, Sloutsky VM. 2004 Is a picture worth a thousand words? The flexible nature of modality dominance in young children. *Child Dev.* **75**, 1850–1870. (doi:10.1111/j.1467-8624.2004.00821.x)

45. Sloutsky VM, Napolitano AC. 2003 Is a picture worth a thousand words? Preference for auditory modality in young children. *Soc. Res. Child Dev.* **74**, 822–833. (doi:10.1111/1467-8624.00570)

46. Robinson CW, Sloutsky VM. 2007 Visual processing speed: effects of auditory input on visual processing. *Dev. Sci.* **10**, 734–740. (doi:10.1111/j.1467-7687.2007.00627.x)

47. Estes KG, Evans JL, Alibali MW, Saffran JR. 2007 Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychol. Sci.* **18**, 254–260. (doi:10.1111/j.1467-9280.2007.01885.x)

48. Hay JF, Pelucchi B, Graf Estes K, Saffran JR. 2011 Linking sounds to meanings: infant statistical learning in a natural language. *Cogn. Psychol.* **63**, 93–106. (doi:10.1016/j.cogpsych.2011.06.002)

49. Echols CH, Newport EL. 1992 The role of stress and position in determining first words. *Source Lang. Acquis.* **2**, 189–220. (doi:10.1207/s15327817la0203_1)

50. Nurmsoo E, Bloom P. 2008 Preschoolers' perspective taking in word learning: do they blindly follow eye gaze? *Psychol. Sci.* **19**, 211–215. (doi:10.1111/j.1467-9280.2008.02069.x)

51. Shukla M, White KS, Aslin RN. 2011 Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proc. Natl Acad. Sci. USA* **108**, 6038–6043. (doi:10.1073/pnas.1017617108)

52. Aslin R, Woodward J, LaMendola N, Bever TG. 1996 Models of word segmentation in fluent maternal speech to infants. In *Signal to syntax: bootstrapping from speech to grammar in early acquisition* (eds JL Morgan, K Demuth), pp. 117–134. Erlbaum, NJ: Mahwah.

53. Curtin S. 2009 Twelve-month-olds learn novel word-object pairings differing only in stress

pattern. *J. Child Lang.* **36**, 1157–1165. (doi:10.1017/S0305000909009428)

54. Ellis NC. 2006 Selective attention and transfer phenomena in L2 acquisition: contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Appl. Linguist.* **272**, 164–194. (doi:10.1093/applin/aml015)

55. Ma W, Golinkoff RM, Houston DM, Hirsh-Pasek K. 2011 Word learning in infant- and adult-directed speech. *Lang. Learn. Dev.* **7**, 185–201. (doi:10.1080/15475441.2011.579839)

56. Xu F, Tenenbaum JB. 2007 Sensitivity to sampling in Bayesian word learning. *Dev. Sci.* **103**, 288–297. (doi:10.1111/j.1467-7687.2007.00590.x)

57. Lany J, Saffran JR. 2010 From statistics to meaning: infants' acquisition of lexical categories. *Psychol. Sci.* **21**, 284–291. (doi:10.1177/0956797609358570)

58. Filippi P. 2016 Emotional and interactional prosody across animal communication systems: a comparative approach to the emergence of language. *Front. Psychol.* **7**, 1393. (doi:10.3389/fpsyg.2016.01393)

59. Filippi P, Gingras B, Fitch WT. 2014 Pitch enhancement facilitates word learning across visual contexts. *Front. Psychol.* **5**, 453–467. (doi:10.3389/fpsyg.2014.01468)

60. Boersma P. 2002 Praat, a system for doing phonetics by computer. *Glot Int.* **5**, 341–347.

61. Moulines E, Charpentier F. 1990 Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun.* **9**, 453–67.

62. Thiessen ED. 2010 Effects of visual information on adults' and infants' auditory statistical learning. *Cogn. Sci.* **34**, 1093–1106. (doi:10.1111/j.1551-6709.2010.01118.x)

63. Bahrick LE, Lickliter R, Flom R. 2014 Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Curr. Dir. Psychol. Sci.* **13**, 99–102. (doi:10.1111/j.0963-7214.2004.00283.x)

64. Lim SJ, Holt LL. 2011 Learning foreign sounds in an alien world: videogame training improves non-native speech categorization. *Cogn. Sci.* **35**, 1390–1405. (doi:10.1111/j.1551-6709.2011.01192.x)

65. Wade T, Holt LL. 2005 Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *J. Acoust. Soc. Am.* **118**, 2618–2633. (doi:10.1121/1.2011156)

66. Filippi P, Laaha S, Fitch WT. 2017 Data from: Utterance-final position and pitch marking aid word learning in school-age children. Dryad Digital Repository. (doi:10.5061/dryad.gh31g)