# From speech to language

## An alternative corpus account of prosodic highlight in continuous speech
## [口語語流中基於韻律亮點之信息預示, 投射與規劃]

Helen Kai-Yun Chen [陳凱芸] and Chiu-yu Tseng [鄭秋豫]
Tamkang University [淡江大學] | Academia Sinica [中央研究院]

This study proposes a novel exploration of the perceived prosodic highlights in continuous speech, focusing on the alternative function of indexing and projecting information content deployment in the context of continuous speech. Given the assumption that prosodic highlight allocation directly reflects the interlocutors' information content deployment, this study foregrounds the perception-based prominences for indexing both the key information (KEY) and the projector (PJR) that projects the deployment of key and/or focal information. Traditionally prosodic prominences were mostly associated with key information. However, the function of prosodic highlight-prompted PJR has been less explored. Two information content planning units (PJR plus its respective projection PJN and KEY) prompted by prosodic highlights were thus established, based on quantitative analyses and discriminative acoustic features. Additional analyses confirm a general *heavy-to-light* information distribution across both units, showcasing that the relative projection trajectory size in the PJR-PJN unit is positively correlated to the position of projection within discourse-prosodic units. Current empirical results, therefore, directly substantiate the cognitive explanation of projection trajectories in speech, as evidence beyond syntactic relationships are drawn and prosodic projection is shown to involve perceived prosodic highlight allocation and information deployment in a fixed pattern. Explorations of prosody-prompted projection creates a venue for a more comprehensive account of the mechanism behind information planning in continuous speech. This allows our findings to facilitate a deeper understanding of the composition of global context prosody and the derivation of linguistic invariants from speech signals.

**Keywords:** prosodic highlight prompting, prosodic projection, information planning and allocation, continuous speech prosody, context prosody

**關鍵詞：** 口語韻律亮點預示、口語韻律投射、信息規劃與配置、
口語語流韻律、語境韻律

## 1. Introduction

This study sets forth an unconventional approach to prosodic highlights that are perceived in continuous speech. Specifically, it suggests an alternative perspective on perceived prosodic highlights annotated consistently in speech that explains their role in relation to information content allocation and planning. Our primary goal is to offer a comprehensive account of the relationship between speech signals in their surface realizations and the underlying representations of languages, and in particular how we can derive regular prosodic patterns from highly variant realizations of speech outputs. Hence, the process of deriving the linguistic invariants from the surface variations of speech signals is crucial. From a perception viewpoint, the more prominently perceived portions of a string of continuous speech signals are traditionally assumed to be associated with focal information and/or new information. This viewpoint is deeply rooted in discussions on information structure (e.g., Chafe 1994, Lambrecht 1994), especially concerning the prosodic marking of discourse referents and reference tracking. Therefore, other studies have suggested or incorporated a straightforward dichotomy of accentuation versus de-accentuation between new and given information (e.g., Halliday 1967, Pierrehumbert & Hirschberg 1990) when approaching information structure. However, we do not fully adhere to the absolute association between accentuation and new information in this instance (i.e., see discussions in Couper-Kuhlen 1986), nor is the tracking of discourse references by the default dichotomy between new and given information our main concern. Instead, our starting point is the prosodic highlights that have been perceived more prominently[1] from continuous speech signals, and their functions regarding global information allocation and planning. We are aware that speakers tend to place prominence directly to correspond to the new/focal information or theme; nevertheless, the following examples also call for further attention, especially regarding speech and interaction:

(1) Nancy: *Jeff* made en asparagus pie.
It was *s: : so: goo:d.*
Tasha: I love it. (Goodwin 1996: 372)

1. We use "prosodic highlight" and "prominence" interchangeably in referring to the same concept throughout the rest of the text.

This draws us to prosodic highlights that are placed not merely on focal/new information but are incorporated to index "specific parts of discourse" (e.g., Falk 2014: 8): For instance, in (1) the enhanced intonation that is placed on the intensifier *so* does not function merely to mark new information. According to Goodwin (1996), the prominently pronounced adverb *so* can be interpreted as a *projector* for the next bit of interaction that *projects* the following adjective of *good* as the main predicate, providing focal and possibly new information. In other words, this study focuses on cases where the prosodic highlights have been placed on strings of speech signals that do not correspond exclusively to new/focal information, nor do they correspond exclusively to the theme. Instead, they function to orient listeners' attention to information allocation under the global speech context.

Therefore, by considering cases such as (1) our goal is to offer a more comprehensive account of the allocation of perceived prosodic highlights in speech via global prosody, particularly for information content deployment. In turn, our approach of associating the function of *projection* with perceived prosodic highlights offers an unconventional account of how to treat emphasized segments that are distinctively perceived from speech signals. We believe the key to effective discourse signal processing lies in capturing the underlying patterns of seemingly random prosodic variations from continuous speech in their surface realizations. Eventually relative information deployment (both within and across phrasal levels) can be shown to be significantly patterned, forming invariant representations and possible categorical distinctions. Thus our eventual goal is to demonstrate that the proposed approach facilitates a more complete understanding of the perceiving and processing of speech prosody within a global context.

As will be demonstrated, we are inspired to explore the function of prosodic highlights for indexing and projecting information allocation because many emphasis tokens (ETs) are now identified as associating with this projection function, rather than directly marking new information. Such findings reflect the importance of the projection function of prominence and call attention to its correlation with information deployment in speech. Moreover, since our analyses are based on prosodic highlights in composing global context prosody, we did not rely solely on the pre-identified morpho-syntactic category as the basic unit for analyzing information deployment. We conclusively identified the information allocation beyond the linear relationship bounded by clausal and sentential concatenations. Eventually the information deployment was explored through an alternative view of a prosodic hierarchy that follows from discourse relations and paragraph associations.

## 2. On projection in previous studies

The concept of *projection* has been approached from various standpoints in literature; for example, in syntax, it has been discussed in relation to the *projection principle* (e.g., Chomsky 1986, Haegeman 1994). In opposition to arguments from a more theoretical viewpoint, *projection* has also been explored from the interaction perspective, such as Auer (2005, 2009, 2015), which focuses on the mechanism of *projection* in interaction and in grammar. Although it was framed by the interaction-based approach, Auer's discussions have been oriented toward a syntax-based explanation of *projection*. Nonetheless, the major difference, lies in the fact that the projection principle of formal syntax focuses on syntactic projection by the subcategorization properties of lexical items; in contrast, Auer suggests that projection depends on a syntactic hierarchy[2] rather than simply linear concatenation (Auer 2009, 2015). Yet both standpoints on *projection* are predominantly syntax- and grammar-based; consequently, the range of *projection* is delineated by either clausal or sentential levels.

---

**2.** To illustrate how projection depends on the syntactic hierarchy, the following example is cited from Auer (2005). In this segment taken from a radio talk show, Caller A has been describing her personal conflict with family members to Psychotherapist B (The original example is in German and here only the sentence illustrating the syntactic projection is presented with the original German text):

(i) 1 B: *Do you have any other relatives or friends*;
*uh who could uhm uh talk about this with you*;
*Or*?
*are you entirely alone against with your mother here in conflict*;
(.8)
6 A: → wissen sie mei GSCHWISter, .h
know 2SG 1SG.POSS brother.and.sister
*you know my brothers and sisters*, .h
7 die halten ALle zu mei MUDda;
3PL stand all by 1SG.POSS mother
*they all stand by my mother*; (Auer 2005: 10–11)

According to Auer, syntactic projection depends on the syntactic hierarchy. To take the internal construction in the turn *meine Geschwister, die halten alle zu meiner Mutter* 'my brothers and sisters, they all stand by my mother' from lines 6 and 7 above, Auer explains that the example demonstrates projections that are allowed by the hierarchical structure, which at least include that 'the Determiner (the first person possessive pronoun *mei*/*meiner*) projects a noun (*Geschwister* 'brothers and sisters'), and the preposition *zu* 'by' projects the noun phrase *meiner Mutter* 'my mother'' (Auer 2005: 12).

Specifically, in interaction-based and conversational-analytic research, the concept of *projection* has been explored in terms of the construction and the organization of turn-taking units. Earlier, Jefferson (1973) suggested that speakers are aware of "possible completion points," and hence, are able to *project* turn endings. Sacks, Schegloff & Jefferson (1974) hold that turn taking depends on subtle features of utterances, mainly syntactic cues, that enable speakers to *project* the end of a turn. The awareness of possible turn completion points was further developed in Lerner's (1996) discussions on jointly constructed turn units. Indeed, the method by which interlocutors become aware of projectability from a currently on-going turn, and how recipients "jump-in" collaboratively to finish co-constructing the turn-so-far provide the most compelling validation of projection in interaction (cf. Auer 1996, 2015, Ford & Thompson 1996). Although most cases of *projection* are associated with features that are interaction-based and turn-related, grammatical resources for turn-internal projections have still been suggested (Huang 2013: 325). Nevertheless, when framed by interaction and turn-taking, cases of projection up until the recognizable completions may well extend beyond sentence boundaries and across turns that are composed of multiple phrases or clauses.

Another vital cue that correlates to the realization of *projection* is prosody. Studies such as Ford & Thompson (1996) have suggested that the projectivity of an utterance can be determined concurrently by its prosody, syntax, and meaning; they demonstrated that both intonation and meaning play a major role in determining the projection of syntactically completed utterances.[3] Other related studies from the conversational-analytic approach, as reviewed in De Ruiter, Mitterer & Enfield (2006), focus on pitch contours at turn finality or turn-yielding cues. As explicated, one of the major drawbacks of focusing on turn-final contours is that these studies often pay attention to features upon projection completion

**3.** In Ford & Thompson (1996) the discussion focuses on turn transitions and their relationship with syntactic, pragmatic, and intonational completions. Based on conversational data, their study examined the convergence of syntactic, pragmatic and intonational completions by the quantitative approach. The following example was used to illustrate how the syntactic and intonational completion converged:

(i) 1 V: She didn't know/ what was going on/ about why they didn't change the knee/.
(Ford & Thompson 1996: 148)

According to the authors, this example demonstrates at least two additional possible syntactic completion points (marked by '/') within the speaker's turn. However, the intonation realizations at the end of the two possible syntactic completion points are relatively level; thus they are not considered intonational completion (Ford & Thompson 1996: 148). In this case the non-completed intonational cue could possibly be taken as an indication of projection of further talk by the current speaker.

alone (De Ruiter, Mitterer & Enfield 2006). However, these features may occur later, and thus cannot account for the anticipated planning from the initiation of projection (De Ruiter, Mitterer & Enfield 2006: 519). After all, listeners do not wait until the point of projection completion to begin the processing of crucial information (Auer 2009).

Studies using an empirical approach have examined projection under various experimental paradigms: while most have incorporated offline judgments by having subjects identify points of projection completion (see De Ruiter, Mitterer & Enfield 2006), other research has also attempted to distinguish if the lexico-syntactic content and intonational cues contribute separately to judgment-making. One study by De Ruiter, Mitterer & Enfield (2006) conducted online experiments incorporating naturally occurring conversational data. Through manipulations that removed either lexico-syntactic or intonational cues, their findings indicate that lexico-syntactic content is a necessary component for projecting points of completion, but intonational cues alone are neither necessary nor sufficient (De Ruiter, Mitterer & Enfield 2006: 531).

Although results of empirical research could single out specific cues that contribute to projection completion and De Ruiter, Mitterer & Enfield's findings (2006) seem to take an opposed position to earlier discussions of projection, we note that in the online processing of interactions, all relevant cues co-exist. While acknowledging the claim by De Ruiter, Mitterer & Enfield (2006) that lexico-syntactic content may have more impact on the decision of projection completion, the present research holds that in online speech processing, intonation, meaning, and syntactic cues are all essential, as argued in Auer (1996). Taking a more holistic view, advance projection plays a crucial role, since the ability to predict upcoming information based on context helps eliminate potential prediction errors that may occur during the communicative process, thus facilitating successful communication (i.e., Clark 2013, Auer 2015, Dilley 2016).

Projection in speech, as noted by Auer, is a *forward*-orientated action that enables interlocutors to make predictions based on an emergent gestalt (Auer 2015: 28). According to Auer (2005, 2015), projection works not merely by linear transitional probabilities in accommodating the next element that is due; instead, it predominantly follows from a hierarchy that facilitates the chance of making correct predictions. From the viewpoint of cognitive processing efficiency, since the emergent structure is anticipated due to the hierarchical relationship, the advantage of projecting upcoming information ahead of time is that it helps reduce the processing load (Auer 2015: 28). After all, when interlocutors can predict upcoming information allocation, it helps them free up the processing load that may be required elsewhere for more complex speech processing (e.g., Auer 2015).

Therefore, *projection* in speech is held as a forward-oriented action that prompts interlocutors to make predictions based on perceived cues, inclusive of prosody. Thus, our research starts with consistently annotated tokens of prosodic highlights in association with the function as the *projector* (PJR) of focal information. We further identify the range of *projection* (PJN) that follows immediately after its respective PJR. Here we consider not only the syntax- and/or semantics-based features, but the **prosody**-related features for the planning of perceptible completion in terms of information allocation. In other words, we do not resort to merely textual analysis for identifying projection and its trajectory; instead, we start from prosodic highlights that are perceived prominently with the possible function as the *projection* initiator. We follow Auer's claim and hold that *projection* itself has a timespan and forms a trajectory (i.e., Auer 2005, 2015). Apart from features that are reflected merely at turn completions, we extend our interest to how information is allocated *throughout* the projection trajectory as reflected by the deployment of prosodic highlights. According to Auer, cognitively speaking, participants may go through a phase of maximal planning during the early part of projection initiation; the processing effort decreases throughout the trajectory (Auer 2005: 9). As will be shown, the results of our analysis directly support this observation, but we use an alternative approach to examine the information-attributed prosodic highlights deployed in continuous speech.

## 3. The present study: A preview

The current study begins from consistently annotated tokens of perceived prosodic highlights with actual emphases (by the discourse-prosodic unit of the *prosodic word*). We attempt to categorize these tokens based on the information content corresponding to each annotated prosodic word with perceived emphases. The initial categorization yields two major information content indexes: The **KEY** index that marks the focal, most salient, and at times new information. Examples include *taifeng* 'typhoon' and *qiwen* 'temperature,' which received emphasis annotation in the simulating weather forecast speech data.[4] The second major index is the *projector* **PJR**, which functions to anticipate upcoming information content allocation. For each case of **PJR**, it is also crucial to identify its range of intended information projection; for example, the *projection* **PJN** trajectory (Auer 2005) that includes at least one piece of soon-to-arrive key information. Using acoustic measurements and quantitative analyses, we were able to establish **KEY** and *projector-projection* (**PJR-PJN**) as two major

4. Please refer to Section 4 for an explanation of the data and methodology.

information content categories that are indexed and prompted by the perceived prosodic highlights.

The two analyses have been set forth to explore information deployment by both information content categories as the planning units and their relative allocation within higher-level discourse-prosodic units (henceforth, **DPU**). Both analyses follow the basic assumption that perceived prosodic highlight allocation in speech is directly associated with information content deployment. Here, our research addresses the following questions: (a) How is information content deployed through the distributed prosodic prompted KEY and *throughout* the trajectory of PJR-PJN unit? (b) How are both information units arranged and planned within the higher discourse-based prosodic levels? As will be shown, the first analysis is devoted to the calculation of *emphasis density*, which is attributed from information allocation. The results establish a fixed pattern of *heavy-to-light* loading across both KEY and PJR-PJN units. In the second analysis, we turn to the location of both planning units by the higher discourse-prosody levels, hypothesizing that their locations, especially for **PJR-PJN**, are directly related to the size of the *projection* trajectory. The results confirm that the longer the *projection* trajectory is, the earlier its **PJR** would be placed within the higher discourse level, beyond the prosodic phrase. Interestingly, the **PJR-PJN** and **KEY** units reflect a *compensatory* relationship with each other by their averaged positions within the higher discourse-based prosody levels.

Therefore, the results foreground the finding that *projection* involves a specific pattern of information planning, which is prosodically prompted and correlated. Ultimately, the validation of a constant pattern of information content planning by perceived prosodic highlights in *projection* has further cognitive significance: We demonstrate that *projection* involves a prediction through information planning that is associated with perceived prosodic prominences beyond syntactically defined units or the linear concatenation of grammatical units. Thus, our findings attempt a novel interpretation (i.e., when compared to the traditional new/given and theme/rheme dichotomies in the discussion of information statues) by offering a more comprehensive justification for information content planning in continuous speech. Finally, the findings regarding prosody-prompted information planning demonstrate that a specific pattern can be derived. This contributes to a more rounded account for the invariants behind context prosody.

## 4. Speech materials and data pre-processing

### 4.1 The speech data

For the current analyses, Mandarin speech data of four genres are incorporated. Two are read speech and the others are spontaneous speech. The read speech include data culled from the Sinica COSPRO corpus (Tseng et al. 2003), covering two genres: (1) speech produced via prose reading tasks (henceforth, **CNA**) by one male and one female native Mandarin speaker; and (2) speech derived from simulating weather forecast tasks (hereafter, **WB**) by one male and one female, both native Mandarin speakers. As for spontaneous speech, one of them is a university classroom lecture in the form of a spontaneous monologue (**SpnL**) taught and delivered by a male professor (Tseng, Lee & Su 2008). The second one is a conversational interaction (**SpnC**) taken from a corpus of Mandarin face-to-face interaction (Chen et al. 2012). The data of a female speaker from one segment of a dyadic interaction have been selected. Table 1 summarizes the total duration and the number of syllables from each genre.[5]

**Table 1.** Summary of total time and number of syllables of the speech data

| Speech genre | Total time (min) | Total number of Syl |
|---|---|---|
| CNA | 50 | 22988 |
| WB | 28 | 14083 |
| SpnL | 145 | 33306 |
| SpnC | 54 | 10756 |

### 4.2 Data pre-processing and annotations

The abovementioned data underwent both automatic and manual pre-processing first, followed by manual annotations for perception-based prosodic information in separated layers. For the pre-processing procedures, speech signals from all the selected data were initially force-aligned into segments, using the HTK Toolkits. The next step involved labor-intensive manual spot-checking by trained transcribers. Thereafter, annotations by experienced taggers were performed in individual layers for the prosody-related information, including the following: (1) levels of DPU, and (2) levels of perceived prosodic highlights.

**5.** Although the quantity of the data does not seem balanced across the different speech genres, we attempted to ensure that there were enough targeting prosodic cues and features to yield valid results.

### 4.2.1 Annotations for discourse-prosodic unit

The key to annotating **DPU** lies in the rationale that prosody-based breaks and boundaries are not constrained by lower word- or phrase-levels, but more crucially, by the necessity of considering higher-levels of breathing- and discourse-associated units. Following the framework of *hierarchical prosodic phrase grouping* (also the **HPG** framework, see Tseng et al. 2005, Tseng & Su 2008, Tseng 2010), five levels of DPUs in hierarchical relationships were annotated across all the speech data. These levels were marked from B1 through B5, corresponding respectively to the following: *syllable* (**SYL**), *prosodic word* (**PW**), *prosodic phrase* (**PPh**), *breath group* (**BG**), and *multiple-phrase speech paragraph* (**PG**). In addition to the lower-level word units and phrasal units, in the **HPG** framework, the *breath group* corresponds by definition to a physio-linguistic unit that is constrained by changes of breath while speaking continuously (cf. Lieberman 1967, Tseng 2010). As for the *multiple phrase speech paragraph* PG, it is identified as the highest level within the hierarchy that is discourse-based and is associated predominantly with topic changes. By default, the boundary breaks, prosodic units, and their relationship within **HPG** can be stated as follows:

(2) SYL/B1 < PW/B2 < PPh/B3 < BG/B4 < PG/B5 (Tseng 2010)

DPUs that are delineated by perceived boundaries and breaks from HPG were manually tagged by experienced annotators in an independent layer. The tagging methods followed from a convention that is similar to the ToBI system (Silverman et al. 1992), in that speech strings were divided into discourse-prosody based units of various sizes. This was conducted by marking boundary breaks in a hierarchical relationship, instead of identifying a singular unit bounded by any type of syntactic relation, one at a time. During and after the annotation process, both intra- and inter- annotator consistency was constantly checked for, to make sure that an agreement was reached.[6]

### 4.2.2 Annotations for perceived prosodic highlight

All the speech data were further tagged manually by trained annotators into perception-based ETs/non-ETs in a separate annotation task. The tagging for perceived prominence was marked by strength levels. They were divided into four relative degrees, from reduced to the most emphasized (e.g., Tseng, Su & Huang 2011, Tseng 2013). The four levels of perceived prominence are defined as follows:

---

**6.** The annotation of DPUs involved at least 10 annotators. As for consistency rate, each annotator had to reach a minimum 80% consistency rate during the initial training before continuing. As for the finalized boundary segmentations, the accuracy had to reach at least 95% of agreement among the annotators.

(3) E0 – reduced pitch, lowered volume, and/or contracted segments
E1 – normal pitch, normal volume and clearly produced segments
E2 – raised pitch, louder volume, irrespective of the speaker's tone of voice
E3 – higher raised pitch, louder volume and with distinctive change of tone of voice

When using this annotation scheme, we specifically stress the fact that only limited numbers of contrastive degrees for prominence could be consistently perceived while processing continuous speech signals online. To annotate perception-based prominence, annotators would simply tag the speech data into a string that consisted of ETs (i.e., E2 and E3) and non-ETs (i.e., E0 and E1) in an independent layer. Note that the identification of ETs/non-ETs is not predetermined by morpho-syntactic units.[7] As explained by Tseng (2013), the purpose of this approach is to facilitate further examination of possible associations between perceived prosodic highlights, with respect to higher-level discourse structures.[8] An additional note is that among the four speech genres, only the spontaneous speech of SpnL and SpnC have been annotated for reduction (E0), as it is assumed that speakers in reading tasks rarely reduce any part of speech production. As for the reliability check, we also consistently verified that the tagging for emphasis levels maintained a consistent performance of over 80% of agreement with both the intra- and inter-annotator reliability checks.[9] Figure 1 provides an example to illustrate the current annotations for both the DPU and prominence levels.

**7.** Since Mandarin does not actually carry pitch accent at the word level, our annotation scheme is distinguished from the model of prosody-related prominence, as proposed by Kohler (1997), or the framework discussed in Baumann, Niebuhr & Schroeter (2016), in that the tagging for prominence levels was not syntactically pre-defined.

**8.** This system for annotating perceived prosodic highlights has been incorporated in several recent studies that explored discourse prosody in continuous speech (e.g., Tseng & Su 2012, Chen, Fang & Tseng 2015), as well as the comparison of prosodic realization in L1/L2 speakers' speech production (e.g., Tseng & Su 2014, Su & Tseng 2015, 2017).

**9.** At least 8 annotators were involved for the annotation of prominence levels. For the annotation process, 1 to 2 "reliable" annotators (who were more sensitive to the prominence level differences) were assigned, and their tagging results served as the "gold standard." The other annotators had to reach an 80% agreement level in their initial training to continue. For the finalized annotation, the accuracy level had to reach at least 95% of agreement among the annotators. Note that after finalizing the annotation, adjustments were still constantly made to fine-tune the annotations so as to better reflect the perception of the speech signals by the prosodic features.

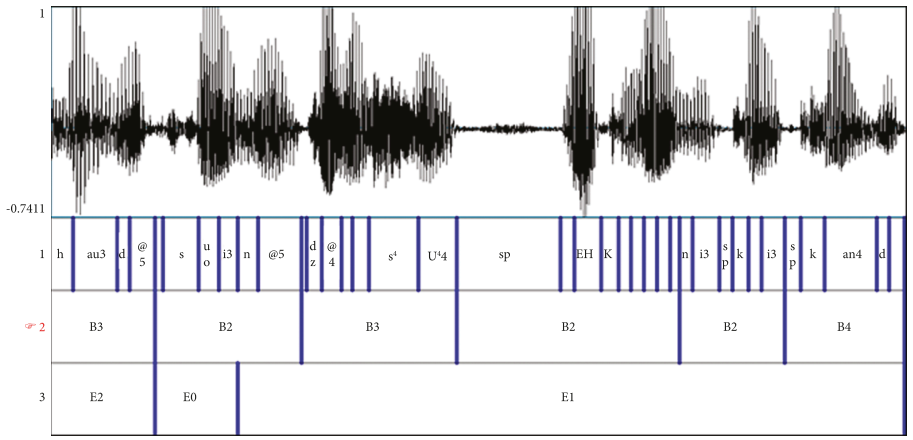



**Figure 1.** Illustration of the annotation schemes for the DPU (in the second layer beneath the spectrogram) and the prominence levels (in the third layer beneath the spectrogram) by using Praat (Boersma & Weenink 2015)

## 4.3 Categorizing prominence prompted information content

After annotating all the speech data with DPUs in a hierarchical relationship and perception-based prominences in relative degrees, we categorized the ETs further by the corresponding information content. We started with tokens that were constantly perceived as more prominent (i.e., E2 and E3) and then broke up each E2 and E3 token by the discourse-prosodic unit at the *prosodic word* (PW/B2) level. The next step involved categorizing the PW with perceivable emphases by its information content.[10] Four major categories were identified: (1) *key information* **KEY**, (2) *projector* **PJR**, (3) *referring expression* **KEY-REF**, and (4) *inferred key* **KEY-INF**.

### 4.3.1 Key information (KEY)

Here, *key information* is the *prosodic word* PW that corresponds to the most salient, focal information or theme in relation to the main topic of the discourse. In general, to associate any PW of prosodic prominence with KEY would be

10. The task of categorization was carried out as a separate annotation task. At least 5 annotators were involved (including the 1st author). The annotators first categorized the PWs that were indexed by prosodic highlight of E2 or E3 into one of the four categories as described in Sections 4.3.1 to 4.3.4. The final annotation had to reach an 80% agreement level among the annotators. Then we turned to each PJR token to identify its projection trajectory. For the trajectory range, we checked and discussed each instance separately to reach a final consensus on the trajectory range for each projection.

like working on keyword spotting in speech processing. Although there were instances where the prosodic-prompted PW corresponded to the newly introduced concept or activated information, this was not a prerequisite condition for identifying **KEY**. In terms of part-of-speech, **KEY** could often be identified through nouns and noun phrases, including proper nouns and all the foreign words. For example, words such as *taifeng* 'typhoon' and *qiwen* 'temperature' that received E2 or E3 tags in the simulating weather forecast data were considered **KEY**. Sometimes PWs of main verbs and predicates could also be considered as **KEY**. Thus *duanci* 'to segment words' from the classroom lecture data would be tagged as **KEY**, as shown in Example (4).

(4) L: Zhongwen shi zhongwen de wenzi shi /**yidui zi**/.[11] Name ni /**bingbu**/
Chinese COP Chinese DE text COP a.CL character then 2SG not
/**zhidao**/ nali shi yige ci. Ni xuiyao-qu /duanci/ cai
know where COP a.CL lexical.word 2SG need.to segment.word then
zhidao-shuo *OK* zhe shi yige sanzici. Zhe shi yige
know ok this COP a.CL three.character.word this COP a.CL
liangzici. Zhe shi yige sizici. Zhe shi yige
disyllabic.word this COP a.CL four.character.word this COP a.CL
danzici. (SpnL)
one.character.word
'(As for) Chinese, the texts in Chinese are presented as **a bunch of characters**. Thus you don't really know which part equals a word. So you need **to segment the words** to know, ok, this is a three-character lexical item; this one a disyllabic word; this one a four-character word and this one a word of one single character.'

In (4), the prosodic highlight indexed noun phrase of *yidui zi* 'a bunch of characters,' and the verb phrase of *duanci* 'to segment words' are both tagged as KEY, based on the main topic of natural language processing from the lecture data.

### 4.3.2 Projector (PJR plus its respective projection PJN)

In addition to placing the prosodic highlight on the focal or salient information, it was noticed that speakers may also incorporate perceived emphasis on a particular PW to head-up the deployment of key information in the coming-up speech production in advance. Following the identification of such a prosodic highlight-

**11.** In this example (as well as the following examples) the word strings in between the slashes have been annotated as PWs in the original data. The PW annotated with E2 or E3 emphasis levels have been indicated by using boldface and underlining it.

prompted *projector* **PJR**, we further delineate the corresponding *projection* **PJN**, whose trajectory covers at least one piece of key information. The annotation of **PJN** also follows from what has been discussed previously in the literature. This covers cases of both *local* and *global* projection (i.e., Lerner 1996, Huang 2013). A local projection can include examples, such as how a numeral plus a classifier project an NP in Mandarin. Other grammatical resources for projection may include instances such as the English example in (1), which demonstrates the intonationally emphasized intensifier when projecting the follow-up assessment in an adjective. However, we have not depended solely on parts-of-speech to identify an emphasized prosodic word as being a **PJR.** Instead, a synthesized decision was reached, based predominantly on perceived emphasis and the corresponding function in projecting the focal information that followed.

The term **PJR-PJN** pair was coined for this purpose. It refers to the prosodic prompted **PJR** that is followed immediately by its respective **PJN.** Two examples of **PJR-PJN** pairs are as follows:

(5) L: Na yeshi /**zuizao de yipian**/ wenzhang. (SpnL)
that also.COP earliest DE a.CL article
'That is also **the earliest entry of** the article.'

(6) L: /**Weisheme** zhi-/ zhijie bidui /**zi ye you**/ kunnan? Yinwei women de
why di- direct match word also have difficulty because 1PL DE
/**ci de**/ jiegou shi feichang *flexible* de. (SpnL)
lexical word DE structure COP quite flexible DE
'**Why** is there difficulty to match **words** directly? (It is) because the composition of the word structure is quite flexible.'

In (5), the prosodic highlight-prompted PW *zuizao de yipian* 'the earliest entry' has been categorized as a **PJR** and has its respective **PJN** trajectory end by the NP *wenzhang* 'article' that follows. As for (6), the prosodic highlight-indexed PW *weishenme* 'why' was tagged as a **PJR**, which entails a projection with its trajectory that extended to the end of the following clause, as initiated by *yinwei* 'because.' In this case, note that the PJN trajectory covers at least one other prosodic highlight indexed PW that was identified as the KEY (there are actually two prosodic prompted KEY: *Zi* and *ci*, which are bolded and underlined). These examples further demonstrate that the prosodic-prompted **PJR** is immediately followed by the corresponding **PJN**. Its trajectory may be of different sizes, from the immediate local (as shown in (5)) to the global ones (as in (6)). Figure 2 presents the annotation scheme of Example (6).

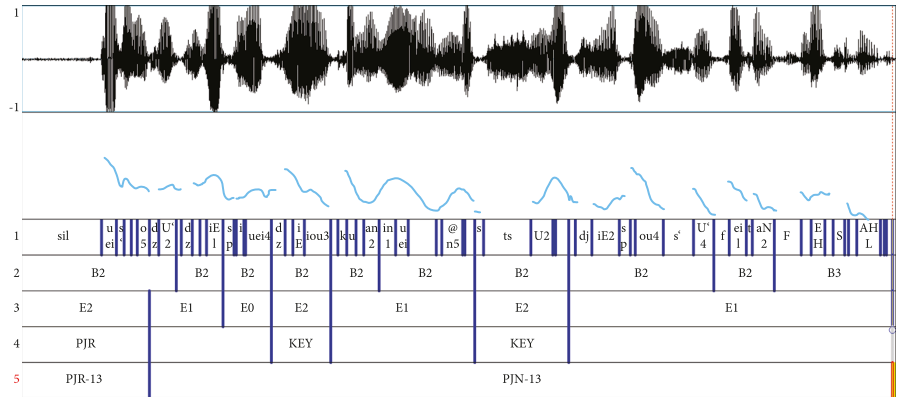



**Figure 2.** Illustration of the annotation schemes for information content categories (in the fourth layer beneath the spectrogram, and instances of PJR-PJN are annotated in a separated layer) by using Praat (Boersma & Weenink 2015)

### 4.3.3 Referring expression (KEY-REF)

The category of *referring expression* was identified when the information content from the prosodic highlight-prompted PW corresponded to a certain referring expression. Here, the referring expression generally covers demonstrative determiners and determinate NPs from the current data. For Example (7), the PW *zheli* 'here,' if tagged with a prosodic highlight, would be categorized as an instance of **KEY-REF**, as it refers to a specific location that is only identifiable through the speech context.

(7) Na yeshi zuizao de yipian zui wanzheng de *paper* suoyi wo ba ta lie
that also.COP earliest DE a.CL most complete DE paper so 1SG BA 3SG list
zai /**zheli**/. (SpnL)
LOC here
'That is also the earliest and the most completed entry of the paper so I list it **here**.'

### 4.3.4 Inferred key information (KEY-INF)

*Inferred key information* is a category that refers to cases when focal information can only be inferred, or when such key information has been mentioned in the context prior to the occurrence of the current prosodic highlight-prompted PW. Some examples are as follows:

(8) Na you yige ren ta yao xie yipian wenzhang guanyu *Taiwan* de
then there a.CL person 3SG want write a.CL article about Taiwan DE
minjian.xinyang tudigong. Ta jiu shangwang qu.zhao tudigong
folk.religion village.diety 3SG then go.on.internet to.find village.diety
jieguo zhao-dao /**yidadui**/. (SpnL)
in.the.end find-PERF a.bunch
'So someone wanted to write an article about the village deity in folk beliefs from Taiwan. And s/he searched for the village deity on the internet and came up with **a bunch**.'

(9) ... buguo shidu yijing ming.xian /**jiangdi**/ (WB)
but humidity already obvious decrease
'...however, the degree of humidity has obviously been **decreasing**.'

In (8), the PW *yidadui* 'a whole bunch of' has been tagged with the prosodic highlight E3. From the context, it can be inferred that the speaker probably wanted the quantifier to refer to *tudigong* 'the village deity,' which was mentioned in the previous context. Similarly, in (9) the PW *jiangdi* 'to decrease' has an E2 tag. From the context, it can be inferred to mean that *shidu* 'the degree of humidity' has been decreasing. We specifically placed these types of examples into one category, as it is assumed that these instances of PW with perceived emphases should be separated from other instances, such as when the prosodic highlight directly marks the focal information itself.

## 5. Analysis I: The information content category

Following the categorization of information content indexed by the perceived prosodic highlight, in the first analysis we focused on the distribution of each information content category. Thereafter, we carried out an acoustic analysis with two major categories: **KEY** and **PJR**. Also, we examined the **PJR-PJN** pair and its interaction with the discourse-prosodic unit boundaries in order to establish it as an information planning unit.

### 5.1 Distribution of information content categories

Figure 3 presents the distribution of the four information content categories that are prompted by the prosodic highlights across the current speech data. The quantitative analysis revealed that, in total, around 68% to 80% of the annotated prosodic highlight tokens with perceivable emphases were categorized as either **KEY** or **PJR**. This implies that, in terms of the information content categorization,

prosodic highlights that are perceived more prominently in Mandarin speech tend to correspond to the indexes of **KEY** or **PJR**. Surprisingly, between these two indexes, the **PJR** outnumbers the prosodically indexed **KEY**. This was found in three of the speech genres, which amount to all except for WB.

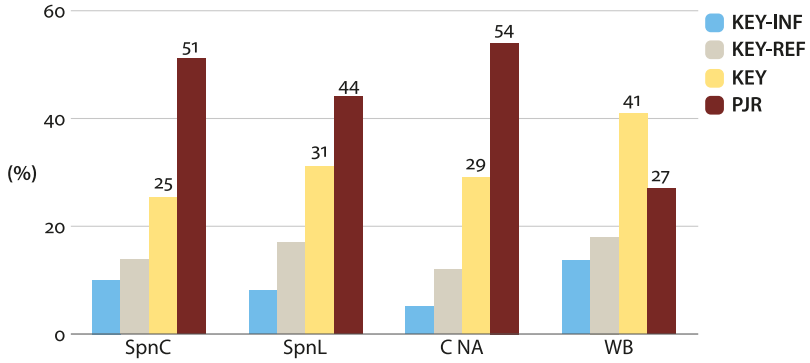



**Figure 3.** The distribution of the four information content categories

### 5.1.1 Discussion

Based on the distribution, we identified ETs in the data as indexes for either focal/key information or the *projector* **PJR** of information allocation within the speech context. To our surprise, **PJR** outnumbers those of **KEY**, which suggests that in continuous speech processing, there is more advanced prosodic prompting than the direct placement of emphases to mark the focal information. As for the exception of the speech genre WB, this finding may indicate that WB differs from the other three genres because it is packed uniquely with far more prosodic highlight-marked key information. As these results point to **KEY** and **PJR** being the two major categories for information content, our following acoustic analysis focuses mainly on these two categories.[12]

## 5.2 Prosodic profiles of KEY and PJR

By focusing on the prosodic highlight-prompted **KEY** and **PJR**, we carried out calculations on acoustic features, including Fo, Fo range, tempo, and intensity from both labels, which were annotated across all the speech genres. Table 2 summarizes the results of averaged acoustic measurements, as derived from the **KEY** and **PJR** tags in read speech (i.e., CNA and WB) and in spontaneous speech (i.e.,

**12.** Due to space constraints, the current paper cannot address the categories of KEY-REF and KEY-INF any further. We hope to discuss both categories in related studies later.

SpnL and SpnC). Table 3 presents averaged acoustic values from the **KEY/PJR** tags across the data.

**Table 2.** Average acoustic measurements from KEY/PJR in read speech (in (a)) and average acoustic measurements from KEY/PJR in spontaneous speech (in (b))*

**a.**

| CNA&WB / Acoustic cues | KEY (mean) | PJR (mean) | *p* |
|---|---|---|---|
| Fo (ST) | 0.035 | **0.407** | <0.001 |
| Fo range (ST) | **0.191** | −0.002 | <0.001 |
| Tempo (syl/sec) | **0.323** | −0.253 | <0.001 |
| Intensity (dB) | **0.022** | −0.203 | <0.001 |

**b.**

| SpnL&SpnC / Acoustic cues | KEY (mean) | PJR (mean) | *p* |
|---|---|---|---|
| Fo (ST) | 0.302 | **0.509** | <0.001 |
| Fo range (ST) | 0.415 | 0.357 | 0.092 |
| Tempo (syl/sec) | **0.639** | −0.053 | <0.001 |
| Intensity (dB) | **0.592** | 0.375 | <0.001 |

* Note that the acoustic measurements reported for Fo in Tables 2 and 3 were based on the unit *Semitone* (ST). Tempo was measured in the standard speaking rate of syllable/second, and intensity was measured in dB. The mean values reported in both tables have undergone standard normalization, the segmental differences (for the tempo measurement) were removed, and the speaker variations were excluded.

**Table 3.** Average acoustic measurements from KEY/PJR across four genres*

| All genres / Acoustic cues | KEY (mean) | PJR (mean) | *p* |
|---|---|---|---|
| Fo (ST) | 0.159 | **0.466** | <0.001 |
| Fo range (ST) | **0.295** | 0.206 | <0.001 |
| Tempo (syl/sec) | **0.469** | 0.135 | <0.001 |
| Intensity (dB) | **0.285** | 0.132 | <0.001 |

* Note that the acoustic measurements reported for Fo in Tables 2 and 3 were based on the unit *Semitone* (ST). Tempo was measured in the standard speaking rate of syllable/second, and intensity was measured in dB. The mean values reported in both tables have undergone standard normalization, the segmental differences (for the tempo measurement) were removed, and the speaker variations were excluded.

### 5.2.1 Discussion

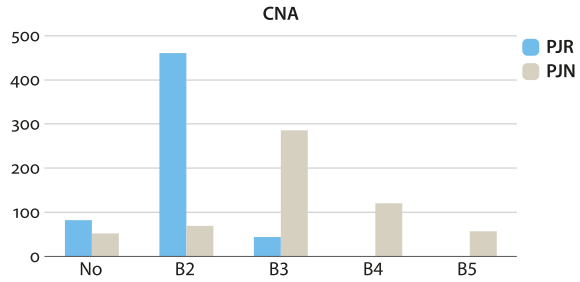
The results presented in Tables (2a) and (2b) show that **KEY** features a larger Fo range, slower tempo and stronger intensity, whereas **PJR** is distinguished only by demonstrating a higher Fo. Note here that all acoustic features could be significantly differentiated between **KEY** and **PJR**, except for the Fo range between the two categories of spontaneous speech. In Table 3, we conducted the calculation again by lumping together all the **KEY** and **PJR** tags across the four speech genres. The results reflect similar findings, in that **KEY** and **PJR** are significantly distinctive because the former tag is realized with relatively larger Fo range, slower tempo, and higher intensity, whereas the latter is realized with higher Fo.

## 5.3 Interim summary

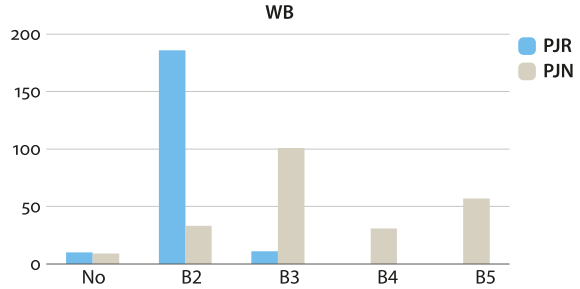
In this regard, we have conducted quantitative analyses, starting from the distribution of the four categories of prosodic highlight-prompted indices. Significantly, we identified **KEY** and **PJR** as the major categories, in terms of distribution. This finding foregrounds the fact that regarding the placement of distinctively perceived prosodic highlights, speakers often incorporate them when directing the recipients' attention to upcoming information planning instead of directly signaling the focal/key information itself. More robust evidence was offered via further analyses of the acoustic measurements extracted from the **KEY** and **PJR** tags. They clearly represent two acoustically distinct categories of speech.

## 5.4 PJR-PJN by the discourse-prosodic unit boundaries

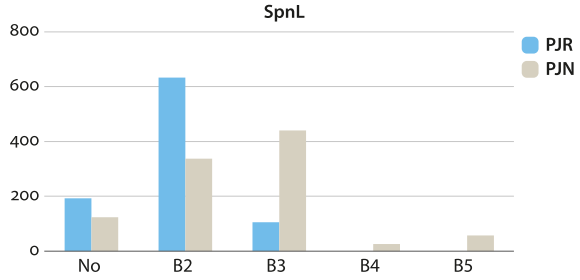
From the above mentioned analyses, our attention was directed to the more frequently occurring prominence-prompted **PJR**. In this section, we turn to the interaction of the information content category and the discourse-prosodic boundaries. As explained in Section 4.3, the categorization of information content was based on the DPU of the *prosodic word* PW. In addition and by definition, the *projection* trajectory of **PJN** to each prosodic-prompted **PJR** can vary between being local or global. Thus, we compared the corresponding DPU boundaries at the end of both **PJR** and **PJN**, which may provide us with clues as to the relationship between **PJR** and **PJN**. The results are summarized in Figure 4. Table 4 presents the **PJR**/**PJN** ending boundaries in relative terms (i.e., if the **PJR** ending boundary is higher/lower/equal to the **PJN** ending, according to the levels of the DPU boundaries).



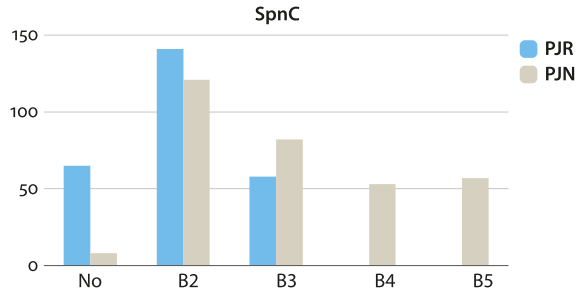

a.




b.




c.




d.

**Figure 4.** The distribution of PJR/PJN ending boundaries by levels of DPUs

**Table 4.** The relative PJR and PJN ending boundaries in comparison

| Speech genre / Boundary ending | Read speech | | Spontaneous speech | |
|---|---|---|---|---|
| | CNA | WB | SpnL | SpnC |
| PJR<PJN | 78% | 81% | 52% | 49% |
| PJR=PJN | 13% | 15% | 35% | 40% |
| PJR>PJN | 8% | 4% | 13% | 14% |

### 5.4.1 Discussion

A definite tendency was observed in Figure 4 for most of the **PJR** endings to fall at the boundary of prosodic word (PW/B2). However, **PJN** tends to fall at a prosodic phrase (PPh/B3).[13] Table 4 shows that around 80% of the read speech and around 50% of the spontaneous data had **PJN** ended by higher boundary levels than the **PJR** ending boundaries. These results suggest that **PJR** is more of a prosodic prompt phenomenon at the local prosodic word level, whereas **PJN** may extend from the word-level upward to the phrasal level, or an even higher discourse level. This implies that **PJR-PJN** pairs are not limited to word-level units. Based on this finding, we suggest taking the prosodic prompted *projector* **PJR** together with its corresponding *projection* as an integrated unit for information planning. Since the *projection* trajectory can be of different sizes, we further examined the range distribution of **PJR-PJN** by the unit of *prosodic phrase* PPh.

**Table 5.** Distribution of PJR-PJN trajectory size by number of PPh

| Genre / PPh # | CNA | WB | SpnL | SpnC |
|---|---|---|---|---|
| 1 | 63% | 77% | 66% | 55% |
| 2 | 25% | 13% | 18% | 28% |
| 3 | 6% | 3% | 7% | 8% |
| Over 3 | 6% | 7% | 9% | 9% |

The above summary shows that around 55–77% of PJR-PJN pairs can be accounted for by up to one PPh. On the other hand, over 90% of projection pairs

**13.** While in the speech data of CNA/WB/SpnL, it is evident that most PJN endings fall at the B3 boundary, there seems to be an exception for SpnC. Although in the SpnC data, most of the PJN endings fall at the B2 boundary, please note that the total cases of PJN ended by B3 and B4 are greater than those by B2.

can be accounted for by one to three PPhs. This illustrates that PJR-PJN pairs are not limited to merely local projection, but also include projections that extend over phrasal boundaries. Given the current results, the **PJR-PJN** pairs together with the prosodic indexed **KEY** are proposed to be established as two major units for information content planning. In the next section, we turn to further analyses that focus on the distribution of information loading across both units, and their location by the higher-level DPUs.

## 6. Analyses by information planning units: Calculation of emphasis density and their position by discourse-prosodic units

Based on the establishment of two major information planning units, namely prosodic highlight prompted **KEY** and **PJR-PJN**, two further analyses are presented in this section. In analysis II, we focus on information loading across both units via the calculation of *emphasis density*. Analysis III explores the correlation between information planning units and their positions within higher-level DPUs. Both analyses are based on the assumption that a direct association can be established between the allocation of prosodic highlights that are perceived from speech signals, and the deployment of information content within both units. Through these analyses, the goal is to provide a comprehensive account for the mechanism behind information content planning in continuous Mandarin speech by analyzing how cognitive loading underpins the planning of information, as well as the compensation between the two information content units by the higher-level DPUs.

### **6.1** Analysis II: Distribution of emphasis density

Analysis II involves the exploration of information planning efforts in terms of cognitive loading during online speech production. Following Auer's view regarding cognitive loading in projection (2005), it is hypothesized that speakers devote maximal planning efforts, starting from the prosodic prompted-projector **PJR**, while the effort decreases gradually throughout the projection trajectory. To substantiate this hypothesis with additional evidence from prosody, we calculated *emphasis density* scores from the distribution of KEY and throughout the PJR-PJN unit. In the following subsections, we first describe the methodology, and then explain the results and discussion.

### 6.1.1 Emphasis density score calculation

The first step of estimating the emphasis density involves merging the reduction E0 tag with E1 from the prosodic highlight annotations for the current spontaneous speech data (i.e., SpnL and SpnC). As mentioned in Section 3.2.2, the read speech was not annotated for reduction (E0). We begin the process by merging the E0 and E1 labels in the spontaneous speech to provide a consistent platform for the calculation of emphasis density scores across the speech genres.[14]

These scores were derived from the scoring system, which is rather *ad hoc* and transparent: Since we assumed that there exists a direct association between levels of perceived emphasis and the information content loading, we simply assigned all the ETs annotated with E1 label a score of 0, with labels E2 and E3 receiving the incremental scores of 1 and 2, respectively.[15] Thereafter, we calculated the *emphasis density* scores from tokens of **PJR**-**PJN** units whose projection extends up to one PPh, while simultaneously estimating the average emphasis density scores from tokens of **KEY** units within this projection range. The calculations of emphasis density scores are based on the unit *prosodic word* PW.[16]

---

**14.** The merging is also based on the general assumption that E1 and E0 are perceptually of minimal distinctiveness.

**15.** The authors would like to note that by using such a direct scoring assignment to equate the most emphasized E3 with the highest score and non-emphasized tokens (E0 & E1) with a 0 score, the methodology may seem rather simple. However, this does not mean that we are not aware of other possible functional associations with prosodic emphases, such as marking contrastive foci. Nevertheless, we believe that the consistent annotation of prosodic highlight levels provided here offer a uniform platform for further comparisons, especially when considering that we are observing data of different speech genres with the goal of more closely exploring the properties of context prosody.

**16.** We adopted the following formula to calculate the emphasis density scores:

ED = Ave (pre_PW_score + current_PW_score + post_PW_score)

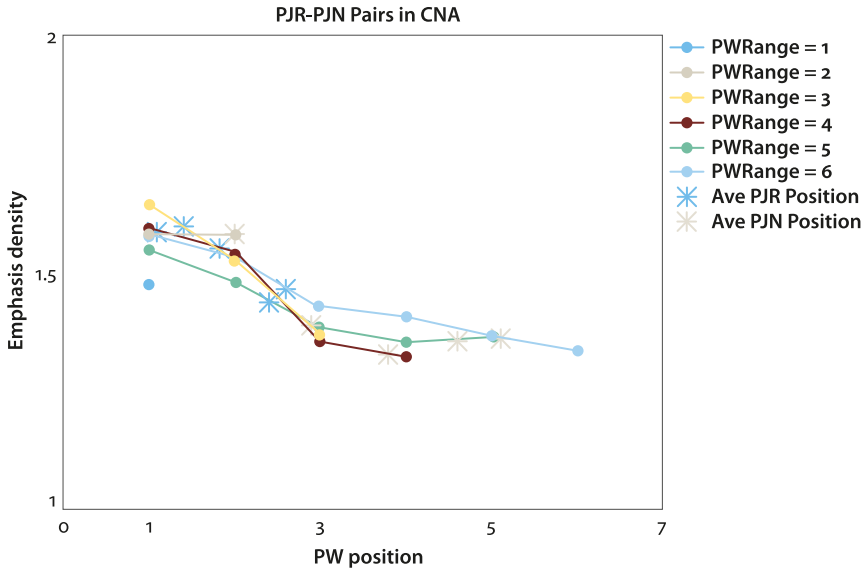
Please note that in calculating the scores, the estimation did not depend solely on the current PW annotated as **KEY** or **PJR**. Instead, the scores were derived by averaging scores from both pre- and post-PWs for that particular **KEY**/**PJR**. By doing so, we were able to consider the emphasis levels annotated for more than just individual **KEY**/**PJR**, but together with the neighboring PWs. Hence, we were able to better capture the distributed information density reflected from the prosodic realization in speech.

### 6.1.2 Results of emphasis density score calculation

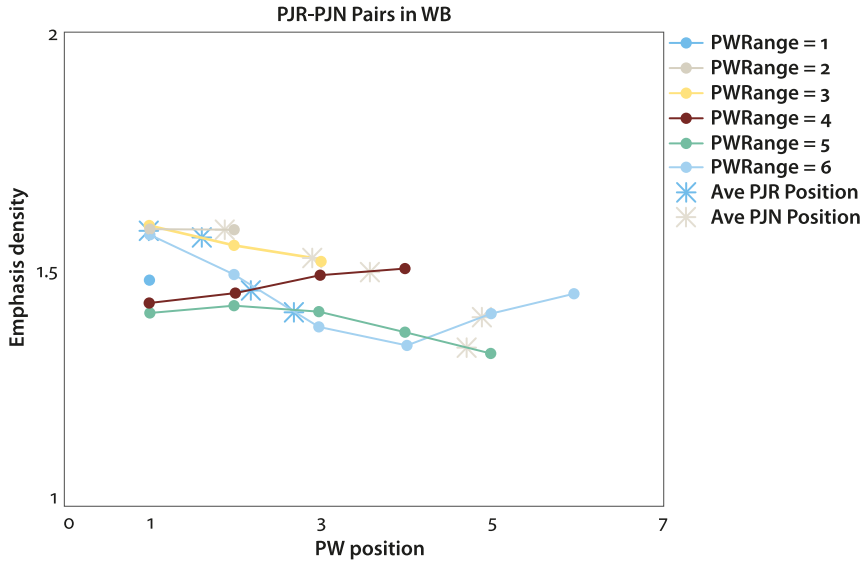
First, the results of the emphasis density scores across **PJR**-**PJN** units are summarized in Figure 5. A general *high-to-low* score distribution can be observed; such that the scores obtained by the end of **PJN** did not rise higher than the scores derived from the beginning of the corresponding **PJR**. Some exceptions are found in the WB data: A higher density score is identified toward the end of **PJN,** where the length of **PJR**-**PJN** equals four PWs. Also, slight rising of the density scores are observed (i.e., less than a 0.1 score difference), toward the end of the **PJN** trajectory, where a unit extends for as long as six PWs. However, in the latter case, the rise never goes higher than the score obtained from the beginning of the corresponding **PJR**. The *high-to-low* emphasis density distribution across **PJR**-**PJN** is otherwise confirmed across most of the other speech genres.[17]

Turning to the density scores from **KEY** (across the four panels in Figure 6) we find that the *high-to-low* scoring generally holds. However, exceptions are observed predominantly in the WB data and in the lecture data SpnL, where the PPh consists of two PWs. Otherwise, the density scores at the end of PPhs never rise to a point higher than the scores from the PPh beginnings. Therefore, it is suggested that the *high-to-low* pattern across the PPh units that contain **KEY** is sustained.

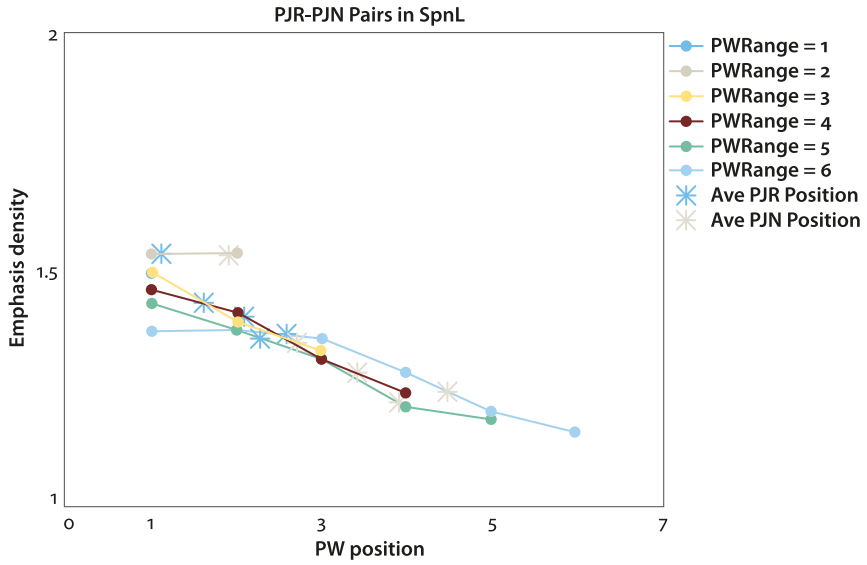
**17.** To validate the current results from the emphasis density score calculation, we also performed a follow-up statistical analysis of the density scores across PJR-PJN pairs and their correlations across the projection range. The additional test was based on PJR-PJN instances ranging from one PPh to three PPhs. The density scores were averaged within each PPh unit to obtain the correlation. The results indicated that there were significant distinctions between PJR-PJN pairs of one PPh and three PPhs for the current read speech (for CNA, $h = 1$, $p < 0.05$; for WB, $h = 1$, $p < 0.05$), as well as for the spontaneous speech of classroom lecture ($h = 1$, $p < 0.05$; but not for SpnC, $p = 0.21$).


PJR-PJN Pairs in CNA
Emphasis density
2
1.5
1
0
1
3
5
7
PW position
PWRange = 1
PWRange = 2
PWRange = 3
PWRange = 4
PWRange = 5
PWRange = 6
Ave PJR Position
Ave PJN Position

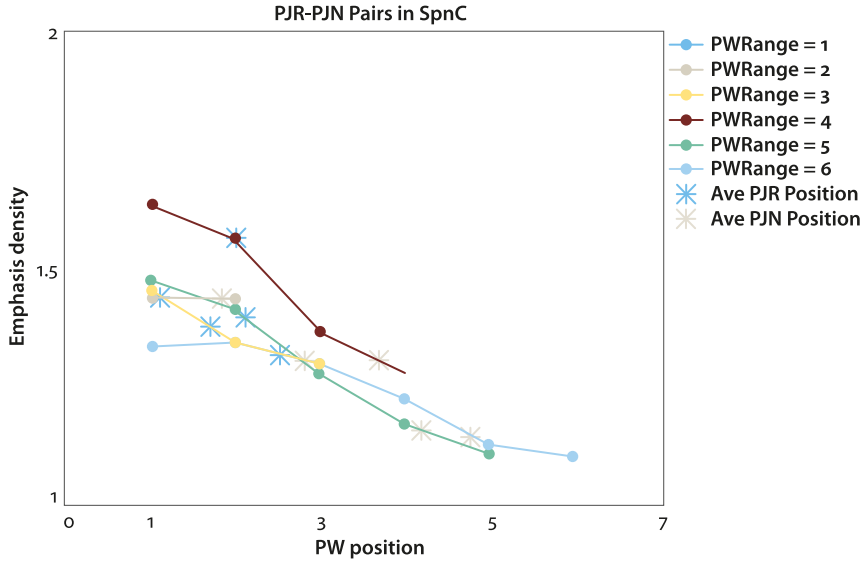
a.


PJR-PJN Pairs in WB
Emphasis density
2
1.5
1
0
1
3
5
7
PW position
PWRange = 1
PWRange = 2
PWRange = 3
PWRange = 4
PWRange = 5
PWRange = 6
Ave PJR Position
Ave PJN Position

b.

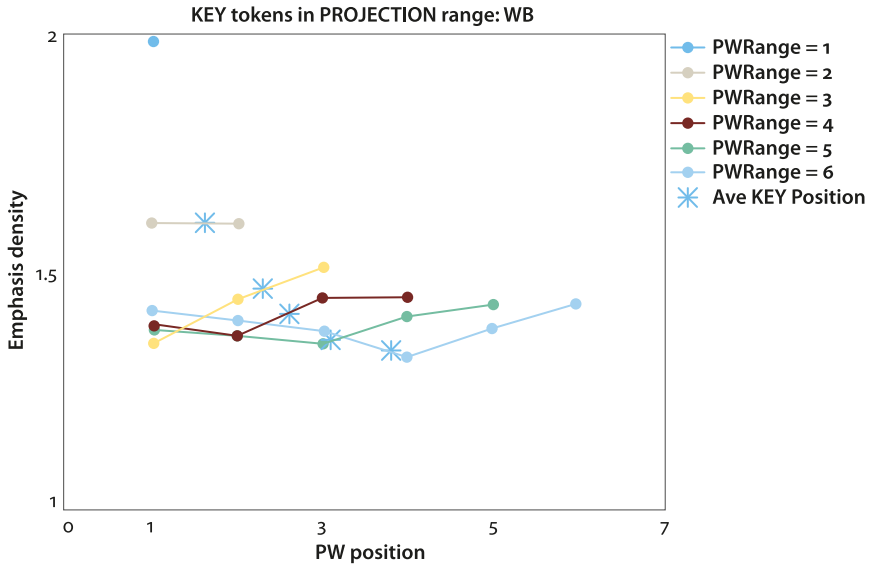



c.

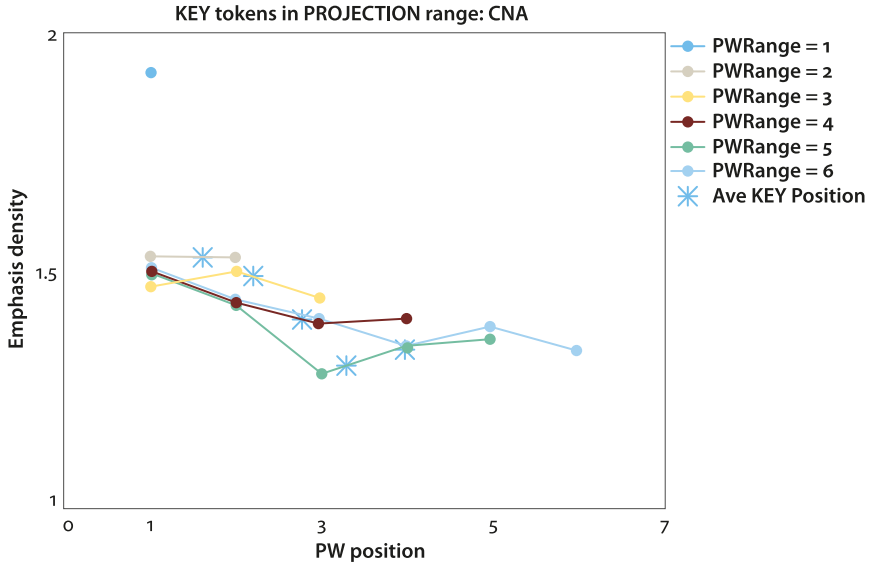



d.

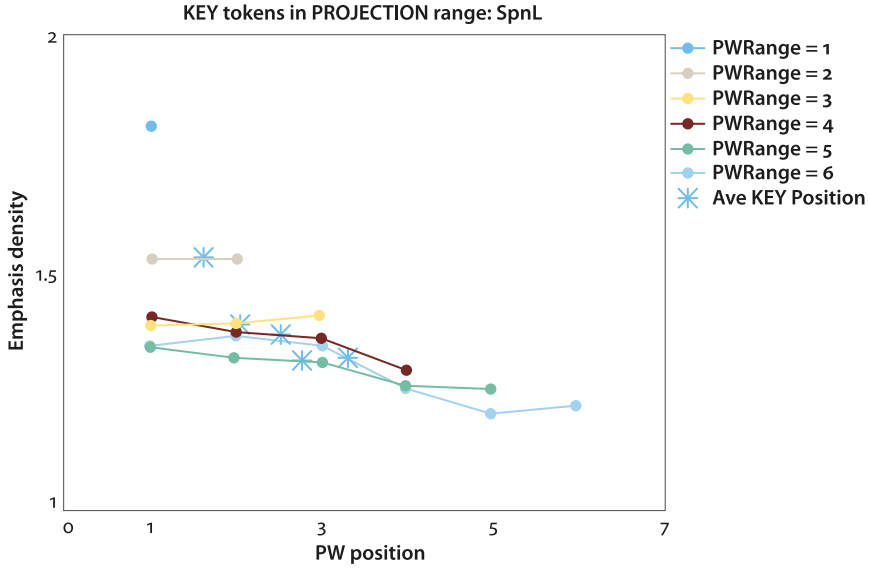
**Figure 5.** Results of emphasis density scores from PJR-PJN units across speech genres (by PW units)


KEY tokens in PROJECTION range: WB
Emphasis density
2
1.5
1
0
1
3
5
7
PW position
PWRange = 1
PWRange = 2
PWRange = 3
PWRange = 4
PWRange = 5
PWRange = 6
Ave KEY Position

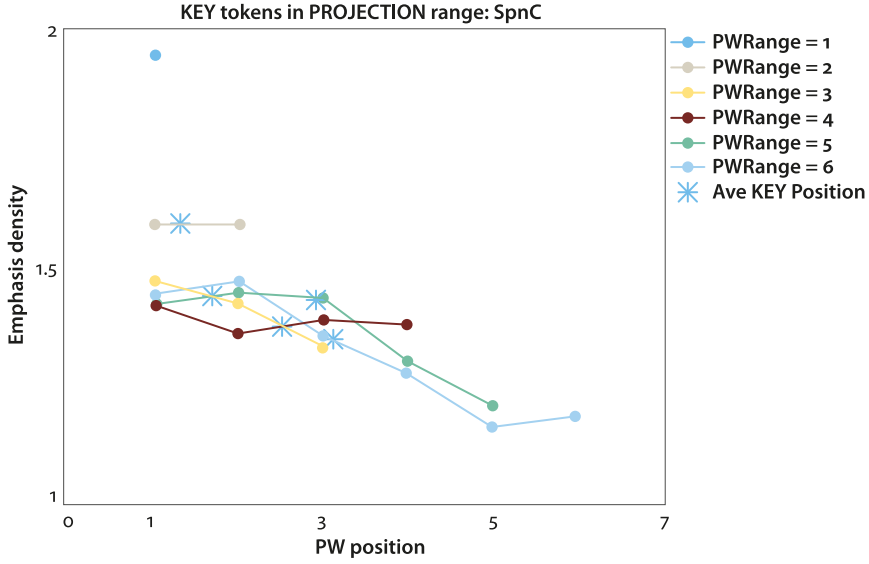
a.


KEY tokens in PROJECTION range: CNA
Emphasis density
2
1.5
1
0
1
3
5
7
PW position
PWRange = 1
PWRange = 2
PWRange = 3
PWRange = 4
PWRange = 5
PWRange = 6
Ave KEY Position

b.




c.




d.

**Figure 6.** Results of emphasis density scores from KEY units across speech genres (by PW units)

### 6.1.3 Discussions

The above results show a general pattern of *high-to-low* emphasis density that is distributed across both information content planning units. This pattern can be translated directly into a tendency of *heavy-to-light* information loading, based on the current assumption. Alternatively, the tendency could reflect that speakers usually start by planning for the heaviest information density at the beginning, after which the information loading is lightened throughout the projection trajectory. This directly substantiates Auer's observation that speakers orient to the maximal planning effort, starting from the *projector* **PJR**, and the cognitive effort decreases throughout the planning of *projection*. Although the *heavy-to-light* information loading may be expected, we are able to provide concrete evidence here via emphasis density scores that are translated directly from perceived prosodic highlights that are allocated across both information planning units.

The few exceptions to the slightly rising density scores throughout both **PJR**-**PJN**/**KEY** information units within the speech data may be due to the differences in the distribution of focal information across the information units, as well as a reflection of genre-specific features. For the read speech WB in particular, we have substantiated earlier that the prosodic highlights are associated with focal information (i.e., Figure 3) predominantly in this genre. A reasonable speculation might be that, in the case of WB, the **KEY** is often distributed toward the end of the projection, and also that this specific genre is packed with far more focal information.

## 6.2 Analysis III: Locating information planning units by discourse-prosodic units

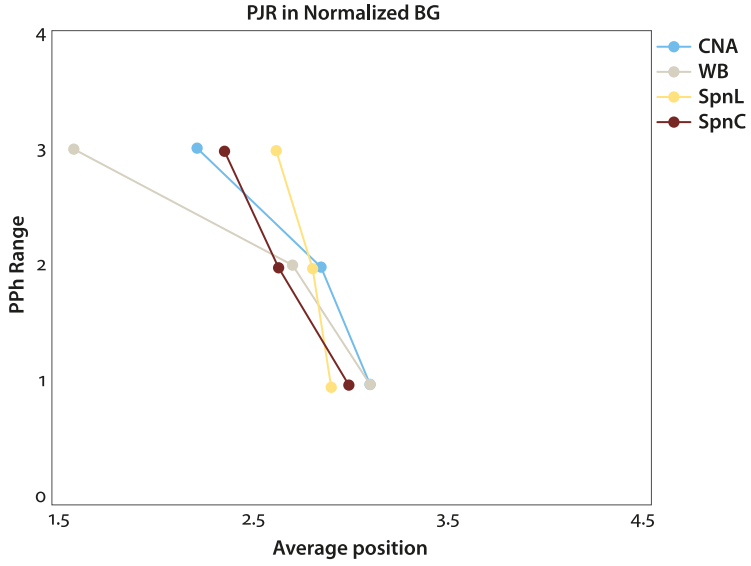
In analysis II above, we calculated emphasis density scores from both information planning units at the *prosodic phrase* level. In the following analysis, while aiming at DPU from higher levels, we examine if the location of both information units might correlate with their planning sizes. In particular, it is hypothesized that the correlation between **PJR**-**PJN** and its relative allocation within *breathing group* BG and *multi-phrase paragraph* PG is a positive one: The larger the *projection*, the earlier the positioned **PJR** would correspond to a later positioned **PJN** ending. Moreover, we examine the locations of **KEY** by BG/PG to further compare the two major information planning units.

### 6.2.1 Estimating the average position of information planning units

To test the hypothesis, we focused on the location of both units within the breathing group unit **BG** and the multi-phrasal paragraph unit **PG** based on the discourse-prosody hierarchy. Given that BG and PG unit sizes can vary drastically across different speech genres, the first step involved a normalization of unit size at the BG/PG levels. We then estimated the average position of the starting point of **PJR** and the end point of **PJN** from each **PJR-PJN** pair, as well as that of **KEY** within the normalized BG/PG. It should be noted that, since the **PJR-PJN** unit can be realized in different trajectory sizes, we have presented the results by the unit of *prosodic phrase* PPh.

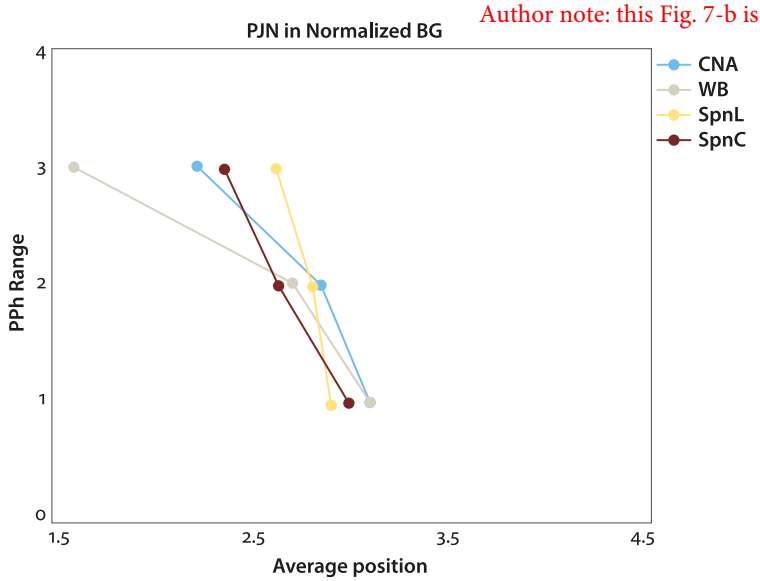
### 6.2.2 Analysis result

As demonstrated in Figure 7, when the **PJR-PJN** unit is composed of two PPhs, the average starting point of **PJR** exhibits a preferred location toward the beginning of the normalized BG. In turn, the longer the **PJN**, the further the trajectory ending will be located toward the end of BG. As a result, a coherent head-tail echo can be observed. Interestingly for the result of **KEY**, it is shown that the information unit is usually located close to the center of BG. This finding indicates that the prosodic prompted **KEY** is distributed evenly across the entire BG, regardless of speech genre.

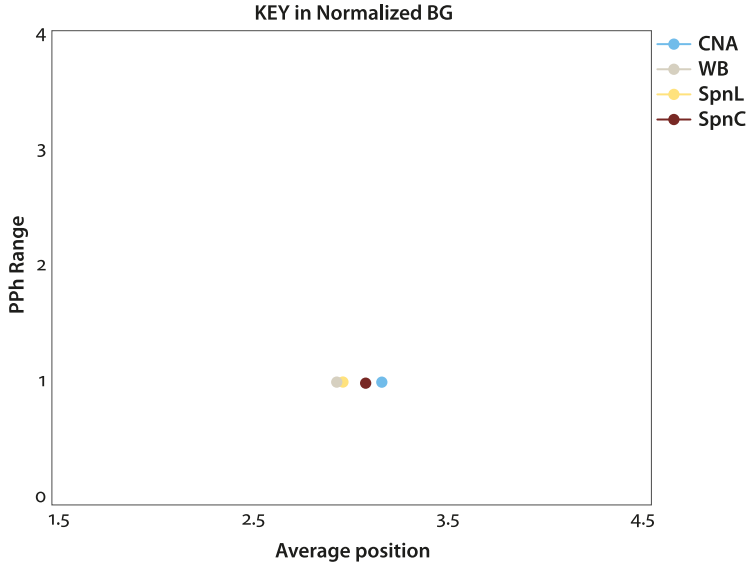



**a.**

Author note: this Fig. 7-b is incorrect.




b.



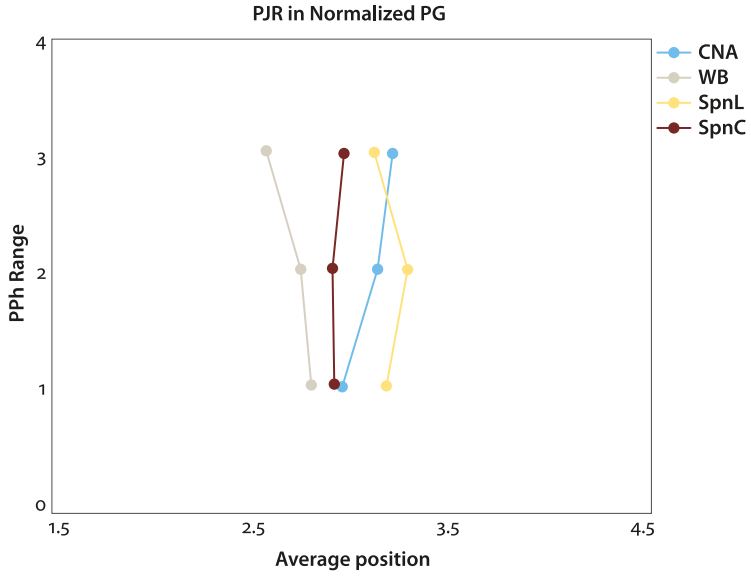

c.

**Figure 7.** Correlation between the information unit sizes and their locations within normalized BG

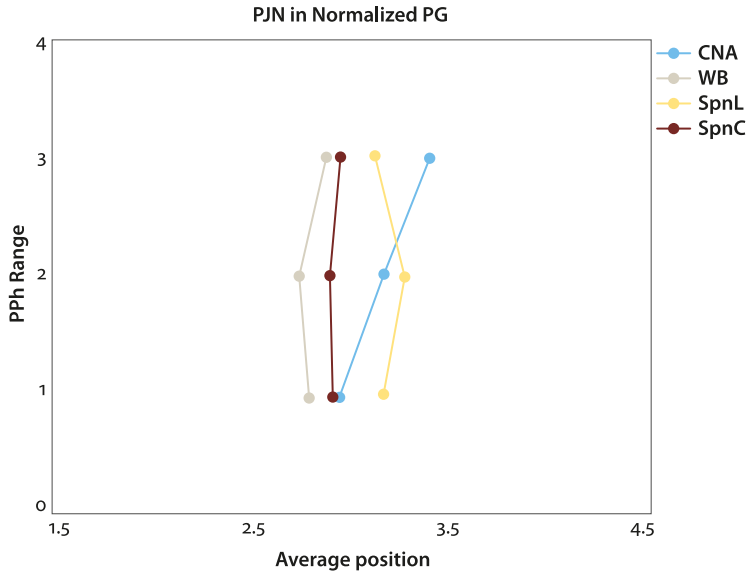
### 6.2.3 Discussions

The above results consequently confirm the proposed hypothesis that a positive correlation can be identified between the location of the information unit and its relative size, especially by **PJR**-**PJN**. In other words, for information planning that starts from the prosodic prompted **PJR** and runs throughout its projection trajectory, the *projection size* turns out to be a crucial factor. When planning for a larger projection trajectory, a speaker would have to arrange for an earlier start to accommodate the projected completion of the entire unit. Interestingly, the planning of the **PJR**-**PJN** unit forms a *compensatory* relationship in comparison with the planning of the prosodic highlight indexed **KEY**: Speakers may plan and signal the key and/or focal information via prominence cues when necessary. Thus, **KEY** can be located at any possible position, so the unit is presented with an average position toward the center of normalized BG.

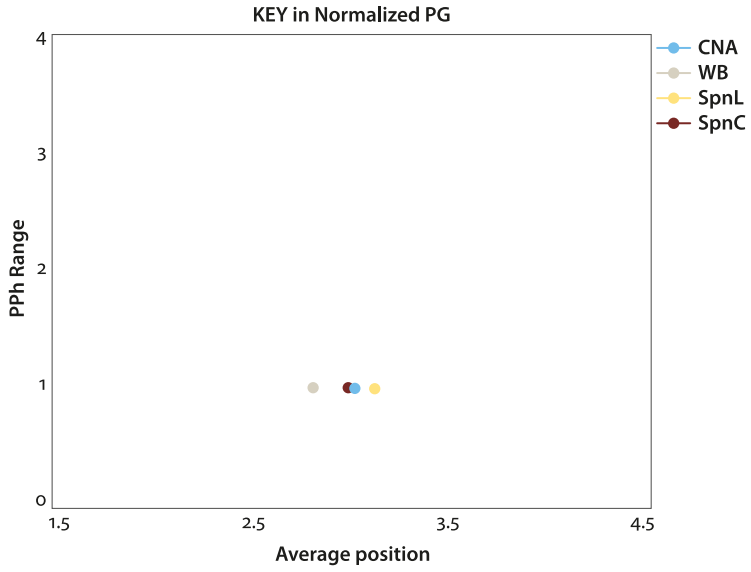
These results lead us to consider whether the same finding will still hold when we turn to the largest discourse-prosodic unit when planning for continuous speeches, namely the *multiple phrase speech paragraph* PG. To verify if this is the case, we examined the correlation between the information units and their positions within the normalized PG. The results are summarized in Figure 8.




**a.**




b.




c.

**Figure 8.** Correlation between the sizes of information planning units and their locations within normalized PG

Results from Figure 8 indicate that the starting points of **PJR** within the normalized PG do not differ distinctively when the projection size increases (i.e., when the projection expands over two PPhs, the **PJR** starting points, on average, did not move significantly away from the PG center). Moreover, the **PJN** ending points demonstrated similar patterns in terms of their locations within PG: The average positions of where the information units start and end remained fairly close to each other toward the PG center. Meanwhile, the averaged positions of **KEY** were also located close to the PG center. These findings imply that both types of information planning units demonstrate an even distribution across PG at the top level from the discourse-prosody hierarchy.

Based on these findings, it is suggested that the planning for both information units is most likely executed at the **BG** level of the current framework in the prosody hierarchy, which consists of discourse-prosody based units that are way beyond sentences. For the significance of this, Tseng, Lee & Su (2008) reported that the average number of syllables in the BG unit in the SpnL data can be around 2.5 times greater than those of the read speech (such as the CNA data), whereas for the highest PG level, the average number of syllables could show an even larger difference. Moreover, focusing on information planning by DPU across speech genres, Chen, Fang & Tseng (2015) demonstrate that speakers plan for a roughly similar size at the *prosodic phrase* (PPh) level from the same discourse-prosody hierarchical framework, but that unit sizes at the higher discourse-prosody levels (i.e., BG and PG) vary drastically. The previous findings further reinforce the fact that information planning in continuous speech across genres does not necessarily match singular sentences perfectly in turn. Instead, the planning of information content at the BG level may well extend beyond clausal boundaries, even going above the concatenation between sentences.

It is noted that the allocation and planning of information at the higher-level DPU is organized and executed most significantly by BG; while speakers may plan for and devote BG to the deployment of information, the top-level PG would possibly be preserved for topic-based discourse planning (such as topic-shifts) and other pragmatic-related processing that is based in speech. Our current findings also contribute to further clarification of the function-wise divisions between BG and PG of the HPG framework, while aspiring to new evidence for information distribution and planning at the discourse levels.

## 7. General discussions and summary

### 7.1 On prosodic highlight-prompted information content units

The current research has been framed by an unconventional approach toward information content planning, focusing on prosodic highlight correlated projection and indexes in continuous speech. It demonstrates that prosodic highlight-prompted projection plays a crucial role in information content allocation for speech, providing vital cues that signal the focal information deployment *in advance*. While identifying the initiation point of projection, our approach is novel in that we started from the prosodic highlights for indexing the projector **PJR**. Thereafter, we delineated the projection trajectory for each prosody-prompted PJR by considering the syntactic and semantic clues in speech together. Thus, a synthesized judgment was reached in defining the projection trajectory, which relied on more than syntactic cues. Furthermore, the current approach attempts to simulate how speakers actually integrate prosodic projection to allow for better predictions on information content allotment. Therefore, instead of concentrating on prosodic cues that are realized only at possible projection completions, we were able to pay attention to how the information is deployed throughout the entire projection trajectory. This approach provided an all-around account for the early projection initiation that are prompted by distinctive prominence cues.

This study first categorized tokens of perceived prosodic highlights by the corresponding information content. We singled out two major categories that are associated with prosodic highlights of actual emphases, namely those marking key information KEY and those indexing projector PJR. The initial annotation was based on speech data of different genres, including both read and spontaneous speech. Not only did we provide evidence that these two categories could be distinguished in the acoustic analyses, but we also found that, surprisingly, PJR outnumbered KEY in most speech genres. Although prosodic highlights have traditionally been associated with focal and salient information, we argue that a more inclusive account is achieved by following the current categorization based on prosodic prominence (i.e., cases where prominence is placed on units like function words). Finally, it was demonstrated through quantitative analyses that more tokens of perceivable prosodic highlights could be incorporated to index speakers' advanced cuing of soon-to-arrive key information, thereby prompting an expectation of the upcoming information content allocation. Thus, it is suggested that prosodic highlights that index the projection initiation in prompting information deployment deserve more attention, as they reflect how speakers incorporate the prominence to co-construct the context prosody.

## 7.2 Toward a cognitive significance for prosodic highlight-prompted projection

Following the categorization of prosodic highlights correlating KEY and PJR, we established two major information content units (with PJR that is immediately followed by its respective projection PJN). Two analyses were conducted, one focusing on the locations of both units within higher DPU levels, and the other exploring the distribution of emphasis density. In terms of location, the two planning units formed a *compensatory* distribution; on average, a center location of the higher-level discourse units for KEY suggests an even distribution. This means that speakers simply place the prosodic highlight on focal information whenever necessary. As for PJR-PJN, its significance surfaces through substantiating that the larger the projection trajectory, the earlier its initiation, and the later its ending will be located within the normalized BG. While it is unsurprising that relatively longer projections may require earlier initiation and later projection endings, our finding alludes to the significance of information planning via prosody in speech. In the end, prosodic projection turns out to be much more patterned and predictable than perceived prominence as directly-marked focal information.

By calculating the emphasis density score, it was revealed that the pattern derived from prosodic highlight projection reflected the significance of information content deployment: The *high-to-low* scoring simply reinforced the fact that speakers tended to prepare for the heaviest information load from the prominence initiated PJR, and such information loading decreased gradually throughout the planning of the projection trajectory. Hence by empirical evidence, we demonstrated how identifiable patterns are generated from prosodic variations in speech signals and eventually driven toward the establishment of meaningful linguistic categories that are based solidly and distinctively on prosodic contrastiveness.

To further account for the cognitive implications, the current results echo Auer's claim (2005, 2015) that high cognitive stress is identified at the beginning of projection. Here, solid evidence is provided by the emphasis density scores across the major information units. Auer's assertion regarding projection was predominantly based on syntax within the hierarchical relationship (2005, 2015). Significantly, the present results suggest that prosodic highlights can also play this role: The allocation of prominence is patterned *throughout* projections. Therefore, projection in speech can go beyond the syntactic arrangements in adjacency and involve perceived prosodic highlights for initiating information deployment. The current study has explicitly illustrated that the planning of information is often executed at a breath group (BG) level in the prosody-

based hierarchy.[18] With the clarification of the level for information deployment through prosody-cued projection, it is suggested that extra processing efforts could be released and possibly re-directed to discourse- and/or pragmatic-oriented planning, which likely takes place at the top level of the same hierarchy.

In sum, the current study contributes to a comprehensive understanding of the mechanism behind information content planning in continuous speech, specifically by the allocation and compensation between two types of prosodic highlight-prompted information content units. We provide solid empirical evidence substantiating that KEY and PJR-PJN form the two major units for information content planning. This alternative approach has been proven to offer a more inclusive explanation than the traditional new-vs-old information structure distinction and even the theme-rheme analysis. Thus, prosodic projection is far more patterned in information content deployment than prosodic highlight-marked focal information, which was shown to be placed randomly. Therefore, we believe that the present study opens avenues for a more holistic understanding of the prosodic constitution of speech output, whereby information structure and planning that goes beyond clausal/sentential structures must also be considered. Future research should substantiate the current annotation scheme for prominence levels and the correlation between perceived prominence and information density by additional evidence that may include the results of online perception experiments. We hope that the current investigation will eventually contribute to an understanding of *context prosody* and advances the account for the establishment of prosody-based linguistic invariants.

## Acknowledgements

The authors gratefully acknowledge the assistance of Mr. Yen-Hsing Chen, Mr. Wei-te Fang, and Dr. Chao-yu Su for carrying out empirical analyses and relevant discussions in the current study, as well as research assistants from the Phonetics Lab at the Linguistic Institute, Academic Sinica (2015–2018) for the annotation tasks included in this work. We would also like to thank the anonymous reviewers and the editors for their insightful suggestions and comments on previous versions of the manuscript. All remaining errors, nevertheless, are ours.

## Abbreviations

| | | | |
|---|---|---|---|
| 1SG | 1st person singular pronoun | 1PL | 1st person plural pronoun |
| 2SG | 2nd person singular pronoun | 3PL | 3rd person plural pronoun |
| 3SG | 3rd person singular pronoun | BA | particle *ba* |

**18.** These findings substantiate the importance of planning the physio-constrained change of breath (i.e., breath group, see Lieberman 1967) in a discourse-prosodic-based hierarchy, such as the HPG framework.

CL classifier
COP copula
DE associative/complementizer *de*
LOC localizer
PERF perfective marker
POSS possessive marker

## References


Auer, Peter. 1996. On the prosody and syntax of turn-continuations. *Prosody and Conversation: Interactional Studies*, ed. by Elizabeth Couper-Kuhlen and Margret Selting, 57–100. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511597862.004

Auer, Peter. 2005. Projection in interaction and projection in grammar. *Text & Talk* 25.1:7–36. https://doi.org/10.1515/text.2005.25.1.7

Auer, Peter. 2009. Online syntax: Thoughts on the temporality of spoken language. *Language Sciences* 31.1:1–13. https://doi.org/10.1016/j.langsci.2007.10.004

Auer, Peter. 2015. The temporality of language in interaction: Projection and latency. *Temporality in Interaction*, ed. by Arnulf Deppermann and Susanne Günthner, 27–56. Amsterdam & Philadelphia: John Benjamins. https://doi.org/10.1075/slsi.27.01aue

Baumann, Stefan, Oliver Niebuhr, and Bastian Schroeter. 2016. Acoustic cues to perceived prominence levels: Evidence from German spontaneous speech. *Proceedings of 8th Speech Prosody Conference*, ed. by Jon Barnes, Alejna Brugos, Stefanie Shattuck-Hufnagel and Nanette Veilleux, 711–715. Baixas, France: ISCA Archive. https://doi.org/10.21437/SpeechProsody.2016-146

Boersma, Paul, and David Weenink. 2015. *Praat: Doing phonetics by computer*. Retrieved November 20, 2015, from http://www.praat.org

Chafe, Wallace. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago, IL: University of Chicago Press.

Chen, Kai-Yun, Laurent Prévot, Roxane Bertrand, Béatrice Priego-Valverde, and Philippe Blache. 2012. Toward a Mandarin-French corpus of interactional data. *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogues*, ed. by Sarah Brown-Schmidt, Jonathan Ginzburg and Staffan Larsson, 147–148. Paris, France: SEMDIAL.

Chen, Kai-Yun, Wei-Te Fang, and Chiu-Yu Tseng. 2015. Information content, weighting and distribution in continuous speech prosody–A cross-genre comparison. Paper presented at the 2015 International Conference Oriental COCOSDA, Shanghai Jiao Tong University, Shanghai.

Chomsky, Noam. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Westport, CT: Greenwood Publishing Group.

Clark, Andy. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36.3:181–204. https://doi.org/10.1017/S0140525X12000477

Couper-Kuhlen, Elizabeth. 1986. *An Introduction to English Prosody*. London: Edward Arnold.

De Ruiter, Jan-Peter, Holger Mitterer, and Nick J. Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language* 82.3:515–535. https://doi.org/10.1353/lan.2006.0130

Dilley, Laura. 2016. Rhythm, context effects, and prediction. Paper presented at the 8th Speech Prosody Conference (SP 2016), Boston University, Boston.

Falk, Simone. 2014. On the notion of salience in spoken discourse–Prominence cues shaping discourse structure and comprehension. *TIPA: Travaux Interdisciplinaires Sur La Parole et Le Langage* [Interdisciplinary Works on Speech and Language] 30:1–23. https://doi.org/10.4000/tipa.1303

Ford, Cecilia E., and Sandra A. Thompson. 1996. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. *Interaction and Grammar*, ed. by Elinor Ochs, Emanuel A. Schegloff and Sandra A. Thompson, 135–184. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511620874.003

Goodwin, Charles. 1996. Transparent vision. *Interaction and Grammar*, ed. by Elinor Ochs, Emanuel A. Schegloff and Sandra A. Thompson, 370–404. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511620874.008

Haegeman, Lillian. 1994. *Introduction to Government and Binding Theory*. Oxford, UK: Blackwell Publishing.

Halliday, Michael Alexander Kirkwood. 1967. Notes on transitivity and theme in English: Part 1. *Journal of Linguistics* 3.1:37–81. https://doi.org/10.1017/S0022226700012949

Huang, Shuanfan. 2013. *Chinese Grammar at Work*. Amsterdam & Philadelphia: John Benjamins. https://doi.org/10.1075/scld.1

Jefferson, Gail. 1973. A case of precision timing in ordinary conversation: Overlapped tag-positioned address terms in closing sequences. *Semiotica* 9.1:47–96. https://doi.org/10.1515/semi.1973.9.1.47

Kohler, Klaus. J. 1997. Modelling prosody in spontaneous speech. *Computing Prosody: Computational Models for Processing Spontaneous Speech*, ed. by Sagisaka Yoshinori, Nick Campbell and Norio Higuchi, 187–210. New York: Springer. https://doi.org/10.1007/978-1-4612-2258-3_13

Lambrecht, Knud. 1994. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511620607

Lerner, Gene. H. 1996. On the "semi-permeable" character of grammatical units in conversation: Conditional entry into the turn space of another speaker. *Interaction and Grammar*, ed. by Elinor Ochs, Emanuel A. Schegloff and Sandra A. Thompson. 238–271. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511620874.005

Lieberman, Philip. 1967. *Intonation, Perception, and Language*. Cambridge, MA: MIT Press.

Pierrehumbert, Janet, and Julia Bell Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. *Intentions in Communication*, ed. by Philip R. Cohen, Jerry L. Morgan, and Martha E. Pollack, 271–311. Cambridge, MA: MIT Press.

Sacks, Harvey, Emanuel. A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50.4:696–735. https://doi.org/10.1353/lan.1974.0010

Silverman, Kim, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert and Julia Hirschberg. 1992. ToBI: A standard for labeling English prosody. *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP 92)*, ed. by John J. Ohala, Terrance M. Nearey, Bruce L. Derwing, Megan M. Hodge and Grace Weibe, 867–870. Alberta, Canada: University of Alberta.

Su, Chao-yu, and Chiu-yu Tseng. 2015. Melody of Mandarin L2 English—When L1 transfer and L2 planning come together. Paper presented at the 2015 International Conference Oriental COCOSDA, Shanghai Jiao Tong University, Shanghai.

Su, Chao-yu, and Chiu-yu Tseng. 2017. How prosodic cues could lead to information center in speech—An alternative to ASR. Paper presented at the 2017 International Conference on Speech Database and Assessments (Oriental COCOSDA), Seoul National University, Seoul. https://doi.org/10.1109/ICSDA.2017.8384443

Tseng, Chiu-yu. 2010. An F0 analysis of discourse construction and global information in realized narrative prosody. *Language and Linguistics* 11.2:183–218.

Tseng, Chiu-yu. 2013. Output prosody—How information highlights are piggybacked by discourse structure. *Zhongguo Yuyin Xuebao* [Chinese Journal of Phonetics] 4:109–124.

Tseng, Chiu-yu, and Chao-yu Su. 2012. Information allocation and prosodic expressiveness in continuous speech: A Mandarin cross genre analysis. Paper presented at the 8th International Symposium on Chinese Spoken Language Processing (ISCSLP 2012), Innocentre, Hong Kong.

Tseng, Chiu-yu, and Chao-yu Su. 2014. L2 discourse and information planning and their prosodic implications. Paper presented at the 2014 International Conference Oriental COCOSDA. Cape Panwa Hotel, Phuket.

Tseng, Chiu-yu, Lin-shan Lee, and Zhao-yu Su. 2008. Spontaneous Mandarin speech prosody-the NTU DSP lecture corpus. Paper presented at the 2008 International Conference Oriental COCOSDA. National Institute of Communication Technology, Kyoto.

Tseng, Chiu-yu, Shao-huang Pin, Yehlin Lee, Hsin-min Wang, and Yong-cheng Chen. 2005. Fluent speech prosody: Framework and modelling. *Speech Communication* 46.3–4:284–309. https://doi.org/10.1016/j.specom.2005.03.015

Tseng, Chiu-yu, Yun-Ching Cheng, Wei-Shan Lee, and Feng-Lan Huang. 2003. Collecting Mandarin speech databases for prosody investigation. Paper presented at the Oriental-COCOSDA Workshop 2003. Sijori Resort Sentosa, Sentosa.

Tseng, Chiu-yu, and Zhao-yu Su. 2008. Discourse prosody and context–Global F0 and tempo modulations. Paper presented at the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008). Brisbane Convention and Exhibition Centre (BCEC), Brisbane.

Tseng, Chiu-yu, Zhao-yu Su, and Chi-feng Huang. 2011. Prosodic highlights in Mandarin continuous speech—Cross-genre attributes and implications. *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, ed. by Piero Cosi, Renato De Mori, Giuseppe Di Fabbrizio and Roberto Pieraccini, 1381–1384. Baixas, France: ISCA Archive.

## Address for correspondence

Helen Kai-Yun Chen
Department of English Language and Culture
Tamkang University (Lanyang Campus)
Yilan, TAIWAN
kaiyun.chen@gms.tku.edu.tw

## Co-author information

Chiu-yu Tseng
Institute of Linguistics
Academia Sinica
Taipei, TAIWAN
cytling@sinica.edu.tw

## Publication history

Date received: 8 October 2020
Date revised: 28 January 2021
Date accepted: 7 June 2021