台灣兒童語言語料庫(TCCM)分詞標準(最後更新時間：3/18/2011)

台灣兒童語言語料庫 (TCCM) 分詞標準是以經濟部 CNS《中文分詞處理原則》為藍本，並參照國內外相關研究、及兒童語言發展的特性做出修訂。

TCCM 的分詞是以能夠產生對應的詞類標記為主要考量，以提升檢索效能、判斷句型複雜度為最終目標。語句中的分詞結果，不一定能與各家的語法分析相吻合，也不能直接作為計算平均語句長度 (MLU) 的單位。

一、分詞原則及層級劃分

1. 分詞原則

規定分詞之依據，包含分詞單位的定義、基本原則與輔助原則。

1.1 分詞單位的定義

具有獨立意義，且扮演固定詞類的字串視為一分詞單位。 根據定義，動詞、名詞、副詞、定詞、數詞、量詞、介詞、地方名詞、連接詞、法相詞、體貌詞、語尾詞皆可依類一一斷開。

1.2 基本原則

中文分詞標準的一般性原則，從語意、語法兩方面來規範，符合語言學理論。

基本原則(1)和(2)配合分詞單位的定義，視同選詞之標準，故為合併原則。

(1) 語意無法由組合成分直接相加而得到之字串應該合為一分詞單位。

凡是組合後意義起變化的字串皆應視為一個詞，意義失去組合性時也都應合為一個詞。

(a) 有些字串因其組合後語意改變皆應視為一個分詞單位。例如:風聲鶴唳(成語) ，撞期、吃醋，三樓(不是三層樓) ，談談、辛辛苦苦、片片、嘻嘻哈哈。

當重疊結構之意義未失組合性，不合併。例如：踢、踢、踢，因該字串之語意是從三個成分組成(重複踢了幾下)。

(b) 合併結構的意義不等於組合意義，故應合併。例如:高中職、中山南北路，依此原則也應該合併為一個詞。

帶專有名詞之合併詞，前方的專有名詞和後方的名詞皆可獨用，意義可以組合成，例如:台北市長等於台北市加市長，但因組合後帶有特定意義，因此仍予以合併。

(2) 詞類無法由組合成分直接得到，應該合為一分詞單位。

(a)字串之語法功能不是從組合而來，應該合併。例如:喝、吃、聽前面加上好構成好喝、好吃、好聽，不能再加賓語，成為不及物動詞，且能被程度副詞“很、十分、非常”修飾，原來的語法特性已經改變。

1.3 輔助原則

輔助原則是彈性運用，視分詞的實際狀況設定分合。輔助原則(1)和(6)為切分原則，(2)~(5)為合併原則。

(1) **分格標記** - 有明顯分隔標記的字串應該切分。

(a) 動賓中插

動賓結構的詞，如:洗澡，可以中插表時態的詞了、著、過，或是中插定量詞。例如:洗 了 一 個 澡。

(b) 述補中插

述補結構的詞，如:打破，可以中插表能力的詞得、不。例如:打得破→打破 得、打不破→打破 不。

(c) 交互中插

兩個詞項可以互相中插。 例如:彎腰和下去，喘氣和過來可以相互交叉形成彎下 腰 去、喘 不 過 氣 來。

(2) **附著語素** - 附著語素和前後詞合為一個分詞單位。

附著語素雖然有獨立意義卻無法獨立扮演一個語法功能，因此和前後的詞合併為一個分詞單位。例如:"演員、救生員…""現代化、合理化…"

(3) **高頻** - 使用頻率高或共現率高的字串視為一個分詞單位。

有些字串因為常常一起出現，結合較緊密，而且中插的情形較少見。因此雖然這些字串中各成分的語意、語法功能未失去組合性，仍合為一個詞。

例1.動詞

(a) 並列結構: 進出、收放、……

(b) 偏正結構: 大笑、改稱、……

(c) 動賓結構: 關門、洗衣、拔草、卸貨、……

例2.名詞

(a) 並列結構: 春夏秋冬、輕重緩急、男女、花草、……

(b) 偏正結構: 象牙、……

例3.副詞: 暫不、既已、不再、……

(4) **雙音節偏正** - 雙音節結構之偏正式動詞視為一個分詞單位。

當一個字串具有動詞的語法功能，符合雙音節結構，且是偏正結構，即可視為一個分詞單位。

例如:緊追其後。 緊追語意、語法功能雖然未失組合性，不含附著語素，也不是常見字串，仍可依此原則合併。

(5) 三音節偏正 - 雙音節加單音節之偏正式名詞視為一個分詞單位。

有些單音節名詞本身可獨立成詞，但是常與前面的雙音節成分結合緊密，故可合為一分詞單位。

例1.“線” 所構成的成分。例如:防衛線、捷運線、木柵線、平均線。

例2.“權”所構成的成分。例如:監護權、領導權、使用權、發言權、優先權。

例3.“車” 所構成的成分。例如:垃圾車、交通車、宣傳車、娃娃車。

例4.“點” 所構成的成分。例如:著眼點、立足點、共同點、爭議點。

(6) 內部結構複雜之詞依不同情況酌以切分。

有些結構雖合併起來過於冗長，若為結合性較強的述補結構或正反問句則不予合併。

(a) 詞組帶接尾詞合併。 例如:花博售票口、紅豆牛奶湯

(b) 專有名稱合併

(i) 專有名詞帶普通名詞。例如: 胡先生、北迴鐵路、二二八事件、和平加油站。動物名或普名與稱謂分開，羊_媽媽、牛_爸爸、郵差_伯伯、太陽_公公

(ii) 書名、歌名及歌詞之專有名詞。例如: 洞內有怪獸 (書名)、海角七號(戲劇名)。

(iii) 特定公司、單位名稱。例如:省自來水公司、師大英文系、中文分詞規範研究計畫。

(c) 正反問句。例如: 喜歡 不 喜歡 、相信 不 相信。

若原本是A不AB的形式，則補足使其完整AB不AB形式。

例如: 喜不喜歡→喜(歡) 不 喜歡、相不相信→相(信) 不 相信

(d) 動詞帶雙音節結果補語則切分。例如:看 清楚、討論 完畢

二、 詞類分合原則

針對不同詞性、不同結構的分合，說明所引用的分詞原則。

1. 動詞

(1)並列結構：洗刷、剪貼

(2) 偏正結構：大笑、靜坐

(3) 主謂結構：心軟、性急

(4) 動賓結構：關門、洗衣、拔草、卸貨

(5) 述補結構：依基本原則(1)(2)一律合併。當補語是結果補語且是雙音節時，依輔助原則(6)切分。中插情形依輔助原則(1)切分。另考慮兒童語法發展的特性，方向補語 到、結果補語 掉 一律分切。

例1.哭濕枕頭、爬上山頭、走 進去、看 清楚、清洗 完畢

例2. V-為: 譯為、流為、批評為、選拔為 〔是述補結構，故合併〕

例3. V-成: 擠成、剪成、歸劃成、堆積成 〔是述補結構，故合併〕

例4. V-作: 鑄作、換作、署名作、轉變作　　〔是述補結構，故合併〕

例5. 述補中插: 打破 得、打破 不　　〔符合輔助原則(1)〕

(6) 重疊結構

若符合基本原則(1) ，則合併。唯中插情形依輔助原則(1)切分。

<1> 嘗試貌

例1. 談談、研究研究　　[符合基本原則(1)]

例2. 說說 看、說 看看　　〔符合輔助原則(6)依定義切分〕

<2> 暫時貌

例: 坐坐 就 走、擦擦 即 可　　〔符合基本原則(1)〕

<3> 程度貌

例: 胖胖的、辛辛苦苦、慢吞吞　　〔符合基本原則(1)〕

<4> 連續重疊詞　　〔不符合任何一項〕

例：「**越吃越胖**」

較為複雜的句型，則改為：

「你_一邊坐_它_有_沒有_一邊叫」

「一邊吃蛋糕_一邊講笑話」

但「越來越」已經成為固定形式，所以切分為「越來越_晚」、「越來越_沒有_同情心」

(7) 正反問句結構

完整形式依輔助原則(6)將之切分，不完整形式則依基本原則(1)、輔助原則(2)切分並補足使其完整。

例1. 喜歡 不 喜歡

例2. 喜(歡) 不 喜歡

(8) 合併結構

依基本原則(1)應合併。唯中插時依輔助原則(1)切分。

例1.上下學、入出境

例2.上、下課，入、出境

(9)中插結構

依輔助原則(1)必須切分。

例:動賓、述補交互中插: 幫 得 上 忙、喘 不 過 氣 來

(10) 附於詞根表時態或特定語法功能。

例如：「去_過」－表過去。

「看_了」－表動作完成。

「燃燒_著」－表動作持續

2. 普通名詞

(1)並列結構

若符合基本原則(1)(2) 、輔助原則(2)(3)四者中的任何一項，則合併。

例:春夏秋冬、輕重緩急、男女、花草

(2)偏正結構

若符合基本原則(1)(2) 、輔助原則(2)(3)任何一項，則合併。

例:象牙、公職人員、財務報表、公共設施

(3)重疊結構

依基本原則(1)應合併。

例:一隻_狗狗、長_痘痘

(4)帶衍生詞綴、接頭/接尾詞

依輔助原則(2)、(5)、(6a)應合併。

例1. 電腦室、業務部

例2. 太空計畫室、國際關係組

(5)簡稱

依基本原則(1)應合併。

例: 男單、女網、空姐、影視、化工、音像

(6)合併結構

依基本原則(1)應合併。

例1.詞頭合併: 高中職、國內外

例2.詞尾合併: 父母親、公私立

例3.套裝合併: 事務局長、台北市長、新竹縣政府

3 專有名詞

依基本原則(1)及輔助原則(6b)應一律合併。

(a)單純詞 - 例: 胡適、台南、布農、貝多芬、光泉

(b)專名+普名(普名是接尾詞) - 例: 阿美族、光復橋、國光號

(c)專名+普名(普名是自由語素) - 例: 胡先生、北迴鐵路、二二八事變

備考: 動物名或普名與稱謂分開，羊_媽媽、牛_爸爸、郵差_伯伯、太陽_公公

(d)縮寫 - 例: 勞基法、消委會、台三線、中常會

(e) 特定公司名稱 - 例: 台北市第一信用合作社、中華電信

(f) 書名、歌名及歌詞 - 例: 鯨魚的生與死(書名)、(戲劇名)

4 定量式

(1)定詞：依分詞定義應予以切分。 唯數詞依基本原則(1)一律合併。

例: 三十五、八萬零二十點七、三又二分之一、百分之四十、三八,000、2・3、20%

(2)量詞：依分詞定義應予以切分。唯重疊結構依基本原則(1)一律合併。

例: 片片、個個

(3)定量詞

依基本分詞原則定詞和量詞應切分。唯重疊結構依基本原則(1)則予以合併。表時間、地點之定量詞依基本原則(1)應合併。

例1. 一 片、一 個　　〔依定義切分〕

例2. 一片片、一個個　　〔符合基本原則(1)其泛指功能〕

例3. 一 片 一 片、一 個 一 個　〔不符合基本原則(1)未具泛指功能〕
例4. 八十四年九月一日三時二十分　〔符合輔助原則(6)〕
例 5. 七十巷二十號之一三樓　〔符合輔助原則(6)〕

5 副詞

唯有符合基本原則(1)(2) 、輔助原則(2)(3)任何一項才予以合併。重疊結構若符合基本原則(1)應予以合併。

例1. 暫不、既已　〔符合輔助原則(3)〕
例2. 不過、要不是、或早或晚　〔符合基本原則(1)〕
例3. 不料、不便　〔符合輔助原則(2)〕
例4. 偷偷、悄悄　〔符合基本原則(1)或輔助原則(2)〕
例5. 叮噹叮噹、砰砰、咻咻咻　〔符合基本原則(1)〕

6 成語、諺語

依基本原則(1)一律合併。

例1. 陰錯陽差、貌合神離、一不做二不休、一而再再而三
例2. 話不投機半句多、虎落平陽被犬欺

三、 各種詞型的分合層級

1. 複合詞

1.1 定量式複合詞

(1)數詞要合併

例:一千八百、百分之三十、三十%、三交二分之一、六十六點五、五成三、七，六五八.四六、AB-8888 、A110048787 、7:20: 3O 、第一、O 二一七八八一三七九九一一五零一、7883799 、2/28-3/31 、三十餘、一百多、二分之一強、四十%以上

(2) 表特定時間、地點之定量詞要合併

例1. 西元 一九九五年 三月 六日 二點 二十分、二十世紀
例2. 八十學年、八十四學年度
例3. 三年五班、五班
例4. 七十巷三十五弄二號之一四樓Bl

(3) 普通定量詞要切分

例: 三 位、五十二 隻、三又二分之一 打、七十餘 位、七十 位 餘、六十多 國、三十來 歲、八 條 半、二十 個 左右

1.2 複合動詞

(1) 並列式

合成雙音節且不可前後互換要合併

例: 醃泡黃瓜、發交相關單位、組建完畢、製播節目

(2) 偏正式

有衍生詞綴、接頭/接尾詞要合併

(a) 偏正式動詞之衍生前綴。 例如:可、好、互、相、自

(b) 偏正式動詞之接頭詞。例如:加、改、重、增、轉、合、代、偷、抽、誤、速、趕、補、複、預、超、回、搶、借、試、大、小、共、對、耐、續

備考:這裡的"共"表"共同"，非"一共"。

(3) 動賓式

辭典已收詞者要合併

(4) 主謂式

辭典已收詞者要合併

(5) 述補式

(a) 方向補語要合併，例如：上來、上來、上去、下來、下去、起來、回來、回去、進來、進去、出來、出去、過來、過去

(b) 結果補語

(i) 補語是單音節要合併

(ii) 補語是雙音節要切分

備考1. 方向補語指的是:上、下、過、起、閱、回、進、出、

2. 起來、下去作時態標記時不和前方動詞合併。

例: 保持 下去、尖叫 起來

3. 起來作評價之用時，和前方動詞合併。

例: **這 件** 衣服 看起來 不錯

(c) 附於詞根表時態或特定語法功能時切分。

例如：「看_過」－表過去。

「看_了」－表動作完成。

「燃燒_著」－表動作持續

(6) V-給、V-到、V-於、V-有、V-為、V-成、V-作、V-掉 另行規定

(a) V-給:要切分

例: 批發 給、寫信 給、分紅 給、取出 給、退回去 給

(b) V-到:要切分

例: 接觸 到、聊 到 半夜、走 到 腿酸、加 到 兩百萬

(c) V-於: 要切分；但動詞是附著詞、或合併後意義改變、或表示比較時要合併

例: 生 於 台北、吝於、有感於、大於、優於

(d) V-有: 要合併)

(e) V-為/成/作: 要合併

(f) V-不得/不了: 要合併

(g) V-掉: 要切分

1.3 偏正式複合名詞

(1)簡單式

(a) 帶語法詞綴要切分

名詞性語法後綴有--們

(b) 帶衍生詞綴要合併

(i) 衍生前綴。例: 老、小、第、阿

(ii) 衍生後綴。例: 氏、某、度、性、家、長、師、員、兒、儿

(c) 2+1 音節

帶接尾詞要合併

(d) 其他

符合下列標準才合併

(i) 帶接頭詞

例: 副校長、準博士

(ii) 含有附著成分

例: 奇案、勇將

(i ii) 語意無組合性

例: 土包子、鐵公雞

(iv) 專指

例: 白菜、黑板

(v) 使用頻率高)

例: 牛肉麵

(2) 複雜式

(a) 簡短、常見式要合併

例: 借書證、租車費

(b) 雖冗長但為專有名詞仍合併

例: 太空計畫室

1.4 複合介詞

(1) 辭典已收錄者要合併

例: 改以

(2) 辭典未收錄要切分

例: 親 至、親 與

1.5 複合副詞

(1) 辭典已收錄者要合併

例: 正在

(2) 辭典未收錄要切分

例: 並 非

1.6 名方式複合詞

(1) 辭典已收錄要合併

2. 專有名詞

(1) 單純詞要合併

例: 胡適、台南、布農、貝多芬、阿爾及利亞、宇宙光

(2) 專名+普名

(a) 普名是接尾詞要合併

例: 阿美族、光復橋、家長會、大漢溪、桃園廠、王董

(b) 與普名合成一專有名詞要合併

例: 胡先生、北迴鐵路、二二八事件、永新加油站

不過動物名或普名與稱謂分開，羊_媽媽、豬_伯伯。

(3) 縮寫要合併

例: 勞基法、奧申委、文建會、臺三線、北二高、中常會

(4) 特定公司名稱要合併

例: 台北市第一信用合作社、台北市捷運公司

(5) 書名、歌名及歌詞要合併

例: 鯨魚的生與死、讓我們看河去

3. 簡稱要合併

例:男單、女網、空姐、影視、化工、音像

4. 合併詞

4.1 無中插

(1) 詞頭合併要合

例: 國內外、高中職

(2) 詞尾合併要合

例: 父母親、公私立

(3) 頭尾合併要合)

例: 中山南北路

(4) 套裝合併

(a) 帶專有名詞之合併詞，組合後帶有特定意義，仍予以合併。

例: 台北市長、正義里長、新竹縣政府

(b) 前面為其他情形要合

例: 事務局長、體育司長

4.2 有中插要切分

5. 重疊詞

5.1 無中插要合併

(1)動詞

(a) 嘗試貌

例: 談談、想想、研究研究、說說 看

(b) 暫時貌

例: 坐坐 就 走、擦擦 就 可

(c) 程度貌

例: 胖胖 的、辛辛苦苦

(2) 名詞。例: 車車、狗狗、小彬彬、痘痘

(3) 量詞。例: 片片、一片片

(4) 擬聲詞。例: 叮叮噹噹、乒乒乓乓

5.2 有中插要切分

例: 說 一 說、想 了 想

6. 正反問句

(1) 完整形式要切分

例: 喜歡 還是 不 喜歡、喜歡 不 喜歡

(2) 不完整形式要補足使其完整並切分

例: 喜(歡) 不 喜歡、漂(亮) 不 漂亮

7. 否定式

不、沒等否定詞均與後接詞切分。

備考: 連用程度高則強制合併，列舉如下，「不錯」、「不過」、「不然」、「不好意思」、「差不多」、「對不起」、「不得了」、「了不起」、「受不了」、「搞不好」、「沒關係」、「沒辦法」、「沒有」。

※「沒有」則與前後詞斷開，例如：「有_沒有」、「沒有_用」

8 中插詞要切分、後移或補足的動作

(1) 動賓中插要切分，例: 洗 了 一 個 澡

(2) 述補中插要切分並後移

(a)三個字的中插，將中插移至最後並斷開。

例如：「看得見」→「Vt|看見_DE|得」

「聽得到」→「Vt|聽到_DE|得」

「看不見」→「Vt|看見_NEG|不」

「聽不到」→「Vt|聽到_NEG|不」

「加什麼油」「跌什麼倒」→「Vi|加油_WH|什麼」「Vi|跌倒_WH|什麼」

「睡一個覺」→「Vi|睡覺_QN|一_M|個」

※ 「吃到飽」的「吃飽」連用性高，因此仍採用中插方式→「吃飽_到」

(b)四個字的中插則直接斷開，不用提取中插字至後。

例如：「套不進去」→「Vt|套_NEG|不_Vi|進去」

「看得出來」→「Vt|看_DE|得_Vi|出來」

(3) 動賓、述補交互中插

(a) 相鄰者優先合併

例: 洗好 澡、剃光 頭、吃飽 飯、彎下 腰 去

(b) 無相鄰者全部切分

例: 幫 得 上 忙、喘 不 過 氣 來

(4) 重疊中插要切分，例: 笑 了 笑、哭 一 哭

(5) 合併中插要補足並切分

例: 初高中→初(中) 高中；國內外→國內 (國)外；

中山南北路→中山南(路) (中山)北路

(6) 正反問句之中插

例：「拿不拿得動」→「拿(動)_不_拿動_得」

「站不站得起來」→「站_不_(起來)_站_得_起來」

9. 成語換字或固定可套換的詞組要合併

兒童語料常見分詞問題

1. 這次 -- 「限定詞+量詞」

-- 例如：「這_隻」、「這_次」、「那_個」...等等斷開。

--「這樣」、「那樣」、「這些」、「那些」、「這裡」、「那裡」均視為一分詞單位。

--「上次」「下次」「每次」若是修飾後面的名詞，則斷開為 |每 次、上_次、下_次。

例如: DET|每 M|次」Nn|上課 都要帶課本。

否則為表時間的副詞類：

ADV|下次 Vt|要 Vt|穿 Vd|給 Nn|阿姨 Vt|看 SFP|喔

Npro:2sg|你 ADV|每次 ADV|都 Vt|要 Vi|筆錄 SFP|啊 !

2. 還是 -- adv.+「是」

-- 若在句首當連接詞時則不切分。例如：「還是」妹妹 比較 聽話。

-- 當連詞作用亦不切分。例如：你 要 吃 麵「還是」飯 ?