香港城市大學　語言資訊科學研究中心
Language Information Sciences Research Centre
City University of Hong Kong

共時語料庫
Synchronous Corpus

切詞規則 Segmentation Guidelines

註 Notes：

1. LIVAC 文本照錄各報章電子版原文，原則上不作改動，由於各地文書處理習慣不同，因此文本中的非漢字字元及標點符號未有統一以單位元字元 (single-byte character) 或雙位元字元 (double-byte character) 表示。

   In principle LIVAC collects the electronic version of newspaper texts in their originals. However, there are different local conventions with respect to non-Chinese characters and punctuation marks, which can therefore appear as single-byte or double-byte characters in the corpus texts.

2. 有關切詞規則第十項（非漢字部份），LIVAC 原有特殊符號標示切詞，現爲符合是次切詞比賽要求的格式，已將該等符號刪除，因此原來視爲一詞的非漢字短句會被切分，唯此改變應不會影響一般切詞系統的運作及結果評估。

   To comply with the formatting requirement of the current bakeoff, the original word delimiters in LIVAC have been removed. Thus some non-Chinese word strings which the following original guidelines treat as one word will now appear as segmented words in the corpus. For example in 10.9 below, <Supreme Governor of the Church of England>, instead of one unit, will now appear as seven separate words. This change, however, should not affect the operation of segmentation systems in general and the assessment of the segmentation results.

   (The following guide is in Chinese.)

規　　則		例
1	名詞 部分	
.1	古今中外各族人名的姓名、譯名概不切分。	張勝利 ｜ 歐陽海 ｜ 江澤民 卡爾・馬克思 ｜ 埃弗斯--威廉斯
	附加「二世」之類的也不切分。	約翰保祿二世
.2	對帶有普遍性或半確定性的稱謂詞，帶稱謂或職稱者，不切分。	李氏 ｜ 王太 ｜ 李伯 ｜ 小張 ｜ 華仔 ｜ 陳x偉 ｜ 陳總 ｜ 張導
	一般性的不合。	陳 妻 ｜ 李 嫌 ｜ 陳 童 ｜ 呂 女 ｜ 林 父 ｜ 陳 犯 ｜ 李 翁
.3	國家全名不切分。	中華人民共和國
.4	民族名、地名中的「族、省、市、州、鄉、區、江、河、山」等不切分。	哈薩克族 北京市 ｜ 浙江省 ｜ 福田區
.5	區域名是雙音節或以上的切分。	香港 地區 ｜ 廣西 自治區
.6	地理名稱不切分。	九龍半島 ｜ 松嫩平原 ｜ 青海高原
.7	村名、街道名全稱不切分。	黃大仙上村 ｜ 皇后大道東
.8	樓宇名稱不切分，	東翠樓
	但由複合詞組成者則切分。	華興 大廈 ｜ 東方 海外 商業 大廈
.9	地名含單音詞的不切分，由複合詞組成者則切分。	中港城 ｜ 黃埔 花園 ｜ 皇后 碼頭
.10	商標名、公司名、機構名由單音詞殿後的不切分; 由複合詞組成的切分。	樂聲牌 ｜ 國泰 AK-47 型 ｜ 彩虹 戰士 號 ｜ 天弓 二 型

	.11	報刊、書籍、影劇、通訊社等專用名稱爲單音詞的合，爲複合詞的切分。	明報 ｜ 聯合 早報 ｜ 金玉集 ｜ 鄧小平 文選 ｜ 法新社 ｜ 新華 通訊社
	.12	“單字名詞+單字方位詞”用來表示地點、機構或時間的，即使是表達本義的，只要有一定的頻率，一概從合。	家裏 ｜ 屋前 ｜ 床上 ｜ 村裏 ｜ 水下 ｜ 河邊 ｜ 黨內 ｜ 選前 ｜ 賽後 ｜ 生前 ｜ 婚後 ｜ 會後
		所表達超出本義或已虛化者合。	手上 ｜ 身上 ｜ 手下 ｜ 根底
		“兩字名詞+單字方位詞”表實義的，一律從分。	學校 裏 ｜ 世界 上
	.13	名詞性的並列結構概不切分。	中英 ｜ 中港台
	.14	經濟名詞或經濟術語不切分。	息率 ｜ 減息 ｜ 護盤 ｜ 跌幅
2	數詞		
	.1	由基數詞及數位詞組成的各種數詞概不切分，數量結構則予切分。	一千八百 ｜ 一九九七 ｜ 34 ｜ 1 億 3502 萬 3 千 ｜ 十七八 歲
	.2	序數詞切分。	第 一 ｜ 第 三十四
	.3	百分數以及含有「數、幾」二字的約數不切分。	百分之三十 ｜ 幾千分之幾 ｜ 1 又 12 分之 1.67 ｜ 數萬 ｜ 十幾二十
	.4	其餘約數切分。	十 來 個 ｜ 三千 餘 人 ｜ 近 三千 多 人
3	時間 詞		
	.1	年、日、時、分、秒切分。	1997 年 ｜ 15 日 ｜ 三 時 廿五 分
	.2	星期、月、日的名稱合。	周一 ｜ 星期日 ｜ 一月 ｜ 八月 ｜ 十二月 ｜ 初一 ｜ 初八 ｜ 年初二 大年初一
	.3	兩個字的合。	上月 ｜ 下周 ｜ 去年 ｜ 前天 ｜ 後年 ｜ 明年 ｜ 周末 ｜ 月初

.4	三個字: A.屬套合詞者不切分。	上月底 | 下周一 | 八月初 | 前年中
	B.非套合詞者應切分	本 月初
	C.含「個」的時間詞應切分。	上 個 月 | 下 個 月 | 半 個 月
	D.「月份」「年度」不切分。	八月份 | 下月份 | 上年度
	E.「來」字應切分。	多 年 來 | 幾 天 來 | 近年 來
	F.其他	較 早 時 | 大前天 | 大清早 | 霎那間 | 一時間
.5	四個字的按以上規則切分。	上 星期三 | 上 星期 末 | 下 世紀 初
4	略語	
.1	簡略語概不切分。	四化 | 亞運會 | 男單 | 臨立會 港人 | 通脹 | 江八條 | 三違反
.2	合併詞不切分。	國內外 | 中小學 | 大中型 | 中低收入
.3	套合詞合。	直升機場 | 心臟病發 | 羽毛球拍 精神病患 | 國民黨籍 | 大連市長 生態學家 | 國防部長
5	二字 結構	
.1	.1.1 在兩個語素中其中有一個語素是粘著的(B)，其結合具有詞匯意義的，有一定使用頻率的，應視爲一個詞而不予切分。	木欄 | 鋼板 | 虎骨 | 雞爪 | 母羊 | 益蛇 | 豆油 | 菜油 | 校門 | 併軌 | 供氣 | 排毒 | 排石 | 設攤 | 貪睡 | 修機 | 節電 | 拔拳 | 搞混 | 刮擦

.1.2 某些粘著的(B)的語素結合面極廣，幾乎都可與另一語素結合成詞，爲免造成此類詞語太多，故只收錄具詞匯意義的詞條或文言詞，餘者不予收錄(在處理上即予切分)。	最高 法院 最後(做連詞用) 你 最好 別管 這 件 事 全 公司 中 張三 身材 最 高 李四 最 矮 他 的 成績 最 好
.2.1 兩個語素均是自由的(F)，而其組合又不具詞匯意義的，或使用頻率不高的，原則上予以切分。	吵 醒 ｜ 走 往 ｜ 飛 往 ｜ 花 香
.2.2a 兩個語素均是自由的，但其組合後引申出新的詞匯意義，即詞義產生了轉義，一加一已不等於二，則不予切分而視之爲詞。	生豬 ｜ 撞破 (有詞匯意義或轉義) 熊膽 ｜ 蛇膽 ｜ 紅花(中藥名)
.2.2b 兩個語素的結合具顯著性，往往表達特別意義，且已相當普遍在使用，此類詞應視爲新詞而予收錄。	上崗 ｜ 下崗 ｜ 到位 ｜ 待崗 ｜ 上證 (大陸新詞) 打壓 ｜ 情治 ｜ 研採 (台灣用語)｜
.2.2c 有不少兩個自由語素組合的雙音詞，結合緊密，使用頻率十分高，已由詞組慢慢演變成詞，則應考慮予以收錄。	停車 ｜ 回國
.2.2d 某些與英文單詞有相對應的雙音詞組合與否，仍應視乎使用頻率，對應因素僅作參考。	牛肉(beef) ｜ 豬肉(pork) 鹿 肉(venison) (使用頻率不高)
.3.1 凡屬文言詞、古語、縮略語，一般不予切分。	叩門 ｜ 易主 ｜ 擒兇 ｜ 拔拳 ｜ 十佳
.4.1 某些動賓關係的雙音詞，在具體語境中會演變成偏正結構的詞語。對這種偏正結構的雙音詞，不予切分。	本 趟 列車 不 掛 車 運行 。 本 趟 列車 共 有 八 卡 掛車 。 沉艦 ｜ 用電 ｜ 用字
.2 並列、主謂、聯動、偏正、動賓、動補等結構:	
A.二個字均爲自由語素、或從屬關係者切分。	寫 信 ｜ 綠 葉

	B.其中一個黏著語素是相當自由的，切分。	本 黨 ｜ 李 某
	C.其中一個或兩個都是黏著語素的合。	手部 ｜ 鞋類 ｜ 逃離 ｜ 打蠟 ｜ 頭痛 ｜ 飛抵 ｜ 打壞 ｜ 受訪 ｜ 做好 ｜ 看成 ｜ 女童 ｜ 力證
	D.結構緊密、使用穩定的合。	回國 ｜ 出院
	E.切分後意思基本不變的切分，切分後意思有變的不切分。	本 國 ｜ 全 黨 綠表 ｜ 白菜
.3	動(形)詞 + 趨向動詞(介):	
	A.下列五種情況不予切分: 1.趨向動詞(介詞)已虛化的，不表示具體方向、方位的，或是表示結果、程度的，或表示引申意的; 2.動詞是黏著語素的; 3.動詞是文言詞的; 4.動詞是方言的; 5.使用穩定、結構緊密的高頻詞。	敢於 ｜ 歸於 ｜ 過於 ｜ 急於 ｜ 樂於 ｜ 達到 ｜ 感到 ｜ 遭到 ｜ 受到 ｜ 超出 ｜ 發出 ｜ 付出 ｜ 退出 ｜ 駁回 ｜ 挽回 ｜ 召回 ｜ 進去 ｜ 看去 ｜ 前去 ｜ 失去 ｜ 揭開 ｜ 解開 ｜ 半開 ｜ 避開 ｜ 岔開 ｜ 落入 ｜ 流入 ｜ 介入 ｜ 併入 ｜ 步入 ｜ 給予 ｜ 賜予 ｜ 賦予 ｜ 寄予 ｜ 交予
	B.下列兩種情況切分: 1.趨向動詞(介詞)有實義的，表示具體方向、方位、地點、時間的; 2.動詞是自由語素的。	定 於 ｜ 生 於 ｜ 用 於 ｜ 走 到 ｜ 帶 到 ｜ 回 到 ｜ 調 到 ｜ 跪 下 ｜ 坐 下 ｜ 彈 出 ｜ 浮 出 ｜ 爬 出 ｜ 流 出 ｜ 調 回 ｜ 退 回 ｜ 收 回 ｜ 帶 來 ｜ 借 來 ｜ 喚 來 ｜ 吹 去 ｜ 走 進 ｜ 衝 進 ｜ 住 進 ｜ 放 進
6	三字 複合詞	
.1	動(形)詞 + 趨(介)切分。	走 過來 ｜ 退 回去 ｜ 優越 於 ｜ 接觸 到
.2	從屬關係切分。	美國 人

.3	接頭詞下列各字合，餘者酌分: 准 總 半 後 禁 無 核 性 零 軟 硬	准新郎 ｜ 總參謀 ｜ 核威嚇 ｜ 性開放 ｜ 零事故 ｜ 軟環境 ｜ 硬道理 超 時代
.4	接尾詞盡量不分，包括: 班 表 兵 波 部 廠 場 車 處 袋 燈 地 點 店 動 度 段 額 法 發 犯 房 費 工 觀 館 光 棍 漢 戶 花 會 火 機 劑 件 界 酒 局 科 庫 礦 力 量 令 樓 率 論 迷 面 民 年 派 片 品 器 氣 切 人 生 師 史 士 式 室 屬 所 稅 獎 司 台 堂 體 天 廳 團 網 尾 物 系 線 箱 形 型 學 炎 業 邑 儀 意 園 院 症 質 種 眾 組	人權法 ｜ 介紹費 ｜ 午飯錢 ｜ 手續費 ｜ 支持票 ｜ 主席台 ｜ 召集人 ｜ 外圍馬 ｜ 伙食費 ｜ 伙食錢 ｜ 休息室 ｜ 光榮感 ｜ 在野黨 ｜ 收錢者 ｜ 住宿費 ｜ 助選團 ｜ 戒嚴令 ｜ 抗議信 ｜ 投注人 ｜ 投注額 ｜ 投訴信 ｜ 私家車 ｜ 和平獎 ｜ 承建商 ｜ 社會課 ｜ 保險界 ｜ 軍政團 ｜ 重案組 ｜ 特別法 ｜ 託運人 ｜ 退休金 ｜ 贊成票
.5	接尾詞「上、下」的用法已超出本義或已虛化者，不予切分;餘者分。(這類結構前面如能加"在"而不影響其接受性者，就需切分。)	實際上 ｜ 事實上 ｜ 大致上 天空 上 ｜ 手續 上
7	四字 詞	
.1	凡屬下列類型的四字結構，不予拆分: a.成語; b.使用穩定的古語; c.四個字不能任意替換的; d.有一定格式的短語。	欣欣向榮 ｜ 愚不可及 ｜ 不爭之論 坐視不理 ｜ 下落不明 ｜ 相持不下 由此可見 ｜ 易燃易爆 ｜ 一心一意 半生不熟 ｜ 大吃一驚 ｜ 猶豫不決
.2	常見的政治術語、慣用語合。	一中一台 ｜ 一國兩制 ｜ 第一時間 一個中國

8 短語		
	超過四字以上的短語概予切分。	
	可兩分可三分的一概兩分。	財政 預算案 ｜ 公園 管理處
	不可兩分的、並列結構的、專名等則三分。	新聞 自由 獎 ｜ 服裝 首飾 展 ｜ 高等 教育 部 ｜ 民主 進步 黨
9 疊詞		
.1	AA 或 AABB 式疊詞合。	看看 ｜ 個個 ｜ 人人 ｜ 明明白白 清清楚楚
.2	ABB 式疊詞合。	綠油油 ｜ 眼睜睜
.3	AAB、ABAB、A 一 A、A 了 A、A 了一 A 式疊詞切分。	說說 看 ｜ 研究 研究 ｜ 想 一 想
.4	AAB=BAA,合	哈哈笑 ｜ 笑哈哈 (但:睜睜眼 ≠ 眼睜睜)
10 非漢 字部份		
.1	所有非漢字符號，包括其它語言的字母、詞串等，保留原有形式，中間不再切分。（註 2）	DOS ｜ H_2O ｜ cc ｜ 4^2 =16 ｜ World News Asia
.2	略語、縮寫語均予收錄。	RAM ｜ SOS ｜ BNO ｜ APEC ｜ 亞太 經合 會議 (APEC)
.3	人名不再切分。（註 2）	Peter Hongda Wu ｜ Franjo Tudjman
.4	地名不再切分。（註 2）	St. Clare Ave. W
.5	機構名稱不再切分。（註 2）	Osaka University

	.6	由英文字母、數目字或「/」、「-」等組成的用以表示型號的詞語不再切分。 但複合詞組仍須切分。	AK-47 | AE-100 | 波音 A2-747 型 飛機
	.7	由英文字母、數目字或「/」、「-」等組成的用以表示序號的詞語不再切分。	A4 版 丨 三 樓 A 座 丨 第 96M 1253 號 丨 車牌 BT4523
	.8	詞串中含有「x」(叉)者，不再切分。	車牌 HK19x7
	.9	一般英語單詞或短句，在文中無特別含意者，視爲一個完整的詞語，中間不再切分。（註 2）	Supreme Governor of the Church of England | gun | abc
	.10	不完整詞語，仍予切分。	澳 (大利亞)
11	其他		
	.1	古語不切分。	是次 丨 甚爲 丨 事發 丨 不果 丨 爲免 丨 問及
	.2	方言不切分。	唔使指意 丨 唔通 丨 玩
	.3	形式相當確定的熟語不切分。	踏破鐵鞋無覓處