附件二

Corpus Encoding Standard (CES)

http://www.mpi.nl/world/ISLE/overview/Overview_CES.html

Last update: 30-8-2000

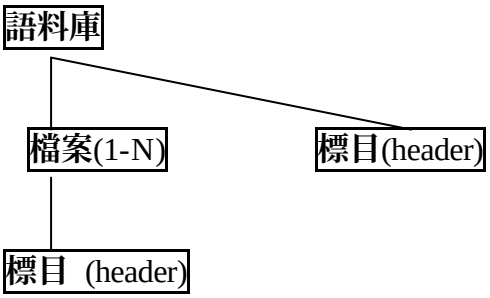
- **簡介**

  Corpus Encoding Standard (CES)**是為了語料庫而發展出的編碼標準**(encoding standard)**，而這些語料庫是語言工程界所使用的。**CES **應用了** SGML**，並符合** TEI **的指引原則。**

- **參考文獻**

  Information about the CES was taken from the CES document version 1.4 (Nancy Ide, 1996)

- **語料庫結構**



- **語料庫資訊**

  藉由 CES **編碼的語料庫包含一個語料庫的目錄**(cesHeader)**和一個或數個檔案**(cesDOC)**。每一個檔案包含有一個文件**(text)**的目錄和文件。另一方面，為了可以指認出次語料庫**(sub-corpora)**，**cesCorpus **的元素都可以重複地套疊**(nested)**，而且每一個元素也可以出現在任何的層次。**

- **文件資訊**

  藉由 cesDoc **定義的檔案包含後接有**<body>**元素或**<group>**元素的一個目錄**(cesHeader)

- **目錄資訊**

  目錄(cesHeader)**提供了有關編碼的電子文件資訊，除了標題**(title)**，作者**(author)**等，也包括編碼相關的資訊。目錄的元素有以下幾大類：**

  - **檔案描述**(fileDesc)**：語料庫或檔案的書目資料描述。是必須的資訊。**
  - **編碼描述**(encodingDesc)**：描述電子文件和其來源間的關係。**
  - **側面描述**(profileDesc)**：提供有關文件各方面描述，尤其著重在使用的語言、建立情況和日期、參與者和其背景、一個文件的描述性分類。**
  - **改版描述**(revisionDesc)**：摘要敘述一個檔案的改版歷史。**

- **後設資料概述**

元素名稱					定義
類型_*屬性*(type*)					目錄所在的檔案類型：屬性「語料庫(CORPUS)」是指目錄是放在語料庫裡；屬性「文件(TEXT)」是指目錄釋放在單一的文件裡。
建立者_*屬性*(creator*)					負責建立目錄的機構
版次_*屬性*(version*)					用來將 CES 編碼的 CES 目錄.elt (CES header.elt)的版次和改版
階段_*屬性*(status*)					目錄的改版狀況：屬性「初版(NEW)」是指目錄的第一版；「最新版(UPDATE)」是指目錄更新後的最新版本
建立日期_*屬性*(date.created*)					目錄內容建立的日期
更新日期_*屬性*(date.updated*)					目錄內容最近更新的日期
檔案描述(fileDesc)					包含有關語料庫和其內文件的書目資料的完整敘述。元素包括有：名稱(titleStmt)、版本說明(editionStmt)、範圍(extent)、出版者說明(publicationStmt)和來源描述(SourceDesc)。標題說明、出版者說明和來源描述是必要的元素。
	標題敘述(titleStmt)				一組有關語料庫標題或個別文件的標題和其組合文件標題的資訊。
		h.標題(h.title)			電子檔案的標題，包含別名或副檔名(subtitle)
		負責人/單位(respStmt)			提供任何負責文件、版本或電子轉寫等人或組織的相關資訊
			工作類型(respType)		以片語描述負責人或單位的實質工作內容
			負責人/單位名稱(respName)		語料庫或文件的出版者/單位，通常是人、地方或機構的專有名稱
	版本(editionStmt)				說明一份文件版本的相關資訊
		版次(version)			
	範圍(extent)				說明電子文件的檔案大小
		字數(wordCount)			說明文件裡的字數

元素名稱					定義
		位元數(byteCount)			說明文件和其標記所佔的位元數
			單位_*屬性*(unit*)		提供位元數計算的單位：屬性「BYTES」-bytes, 數性「KB」-kilobytes, 屬性「MB」-megabytes, 屬性「GB」-gigabytes
		範圍介紹(extNote)			有關語料庫或文件目錄的任何相關資訊
	出版單位(publicationStmt)				有關語料庫及其內部文件的出版或發行的資訊集
		發行者/單位(distributor)			提供發行文件或語料庫的人或機構名稱
		發行者/單位地址(pubAddress)			包含發行者/單位的郵政地址
		電話(telephone)			提供發行文件或語料庫的人/單位的電話號碼, 須符合 ITU-T/CCITT 的形式, 建議參考 E.123
		傳真(fax)			提供發行文件或語料庫的人/單位的傳真號碼, 須符合 ITU-T/CCITT 的形式, 建議參考 E.123
		電子地址(eAeeress)			提供發行文件或語料庫的人/單位的電子地址。可多值著錄, 所以可以是不同類型的地址
			類型_*屬性*(type*)		提供不同類型的電子地址, 例如：電子郵件、網址、ftp 位址等。
		取得權(availability)			提供一份文件的任何取得資訊, 包括使用或發行限制、著作權等。
			領域_*屬性*(region*)		電子文件所採用的權限範圍
			層級_*屬性*(status*)		提供文件現階段取得權限的識別碼
		識別碼(idno)			提供一組號碼(例如 ISBN)以識別書目資料
		出版日期(pubDate)			任何形式表示的出版日期
			日期表示_屬性(value*)		以 ISO 8601 的格式表示日期
	來源描述(sourceDesc)				提供有關生成電子文件的來源紙本的書目相關資料
		書目結構(biblStruct) 元素量(1-N)			包含結構化的書目資料引用, 其中只要出現書目的次元素, 並以特殊的次序排列

元素名稱						**定義**
			可拆式(analytic)			**包含描述一個不是獨立出版品的項目(例如：一篇文章或一首詩)的書目元素；而此項目是發表在專書、刊物內的。**
			專著(monogr)			**包含描述一個獨立出版品(例如：書籍、期刊)的書目元素**
				h.**標題**(h.title)		**作品的標題**
				h.**作者**(h.author)		**作為引用參考文獻之用，包含作品的作者(人或單位)名稱。必須遵守姓在前、名在後的格式**
				負責人/單位(respStmt)		**提供任何負責文件內容、版本或電子轉寫等人或組織的相關資訊**
				版本(edition)		**提供一些文件的詳細版本資訊**
				印行者/單位(imprint)		**和出版或發行書目資料相關的資訊群組**
				識別碼(idno)		**提供用以識別書目資料的標準識別碼(例如** ISBN)
					類型_*屬性*(type*)	**簡寫的名稱(例如：**ISBN**)，用於指出識別碼的類型；除非提供很明確地基值，例如** ISBN
				書目範圍(biblScope)		**用以定義參考文獻的範圍，例如可以是一列頁數、有名稱的部份或一部較大的作品**
					類型_*屬性*(type*)	**用以標示元素傳達的資訊類型，例如：**PP **表示頁數或頁數範圍；**VOL **表示卷數；**ISSUE **表示期數**
				書目介紹(biblNote)		**描述語料庫內或文本目錄內相關的參考書目訊息**
				出版者或單位(publisher)		**人、地方或機構的專有名稱**
					類型_*屬性*(type*)	**有關名稱的類型**(PERSON: **指人的姓名；**PLACE **指地方的名稱；**ORG: **指定期出版刊物的組織**)
				出版日期(pubDate)		**以任何形式表示的日期**
					日期表示_*屬性*(value*)	**以** ISO 8601 **的格式表示日期**

元素名稱					**定義**
				出版地點 (pubPlace)	**書籍、文章的出版地點**
編碼描述(encodingDesc)					**用以描述電子文件及其來源之間的關係或電子文件的來源的檔案**
	計畫描述 (projectDesc)				**詳細描述電子檔案編碼的目的**
	取樣陳述 (samplingDecl)				**以散文的方式描述語料庫內文本取樣的原理和方法**
	編輯陳述(editorialDecl)				**描述當為文件編碼時，所使用的編輯原則與實作**
		一致性 (conformance)			**提供** CES **對文件或語料庫的等級劃分**
			等級_*屬性*(level*)		**給與** CES **一致性的等級，例如：**1, 2 **或** 3
		轉換(transduction)			**描述文件轉換時所依循的原則，可以適用在由錄音帶轉寫成文字形式、或者電子形式間的轉換**
		修正(correction)			**在建立語料庫一個或多個部份時，一組修正程序的應用**
		引用(quotation)			**在編輯時，原始檔內相關引號的應用**
			引號_*屬性*(marks*)		**指出引號在文件理是否被保留作為表示標籤內容**(tag content)**之用。**(NONE: **無引號保留；** SOME: **一些引號保留；**ALL: **所有引號皆保留**)
			形式_*屬性*(form*)		**指出引號在文件裡是如何指示**(STD: **標準化的引號使用，開和閉引號** (open and close quote marks)**是有區別的；** NONSTD: **開和閉引號是沒有區別，皆以**????**表示；**UNKNOWN: **引號的使用未知**)
		連字號(hyphenation)			**簡述在被編碼的版本裡，行尾**(end-of-line)**連字號的處理方式**
		斷詞(segmentation)			**描述文件斷詞的原則，是以句子、聲調單位或圖形層級等**
		標準(normalization)			**應用於建立語料庫的一個或多個部份的標準程序**

元素名稱					**定義**
			方法_*屬性*(method*)		**用來指出標準化是否沒有採用符號或採用編輯的標籤**(TAGS: **表示標準化使用了標籤；** SILENT: **沒有使用任何符號來標準化**)
	標籤陳述(tagsDecl)				**詳細描述** SGML **文件裡所使用的標籤**
		標籤使用方式 (tagUsage) **元素量**(1-N)			**與目錄相關的語料庫或文件裡，使用特殊元素的資訊**
			gi*		**標籤所標示的元素名稱**(一般辨識名稱)
			occurs*		**表示出文件裡元素出現的次數**
			wsd*		**可以使用在**<tagUsage> **元素裡，用以指出在文件裡元素的每次出現皆表示了特殊的字元組**
	參照描述(refsDecl)				**指示出這份文件的正統參照**(canonical references)**是如何建立的**
	類別描述(classDecl)				**包含一系列的**<**類別**(category)>**描述，用以定義語料庫內文件的分類碼**
		分類(taxonomy) **元素量**(1-N)			**用以定義分類文件的類型**
			類型(category)		**包含個別描述類型的詞或一對特徵值**
				類型描述 (catDesc)	**以短文方式描述文件分類裡的一個類型**
側面描述(profileDesc)					**提供一份文件的其他不同資料，包括語言、建立的地點、時間、參與者及其背景、及其分類等**
	建立(creation)				**包含一份文件的來源資訊**
	使用的語言(langUsage)				**包含一組資訊描述有關文件的主要語言、次要語言、登錄者、方言等**
		語言(language) **元素量**(1-N)			**描述一份文件的主要語言、次要語言、登錄者、方言等**

元素名稱					**定義**
			Iso639*		**提供** ISO 639 **裡的標準語言碼，以下列之一種形式呈現**：ISO 639 **裡的二字母碼；**ISO 639-2 **裡的** 3 **字母碼；或上兩種方式之一加上延伸的國碼，此碼是從** ISO 3166 **裡得出**
			類型_*屬性*(type*)		**指出語言的類型，例如：次要語言、方言等**
	字元使用(wsdUsage)				**一組描述文件裡的字元的資訊**
		書寫系統(writing System) **元素量**(1-N)			**一份文件內所使用的字元**
	文件類別(textClass)				**一組描述文件性質或標題的資訊。這個性質或標題是採用** thesaurus **的分類標準**
		類別參照(catRef)			**一些分類學上所定義的一個或多個類別**
			目標_*屬性*(type*)		**用以辨識出文件的類型，主要藉由** IDREF **屬性指到由語料庫目錄所定義的一或多個<類型>元素**
			架構_*屬性*(scheme*)		**用以指出分類的架構**
		h.**關鍵詞**(h.keywords)			**包含一列關鍵詞用來指出一份文件的主題或性質，每一個關鍵詞以一個詞項**(term)**作為標籤。標準列將由** EAGLES/PAROLE **提供**
			關鍵術語(keyTerm) **元素量**(1-N)		**包含專業術語，特別在一列描述關鍵詞上**
	翻譯(translations)				**一組有關已存在的翻譯文件的資訊**
		翻譯(translation) **元素量**(1-N)			**給與有關文件的翻譯資訊。這個標籤需要全球語言的屬性和書寫系統**(wsd)**屬性。**
			翻譯.位置_*屬性*(trans.loc*)		**提供有關翻譯的位置資訊，例如路徑/檔案名稱、**URL **等**
		翻譯者(translator)			**翻譯者的姓名**
	註解(annotations)				**一組和文件相關的註釋檔案資訊**
		註解(annotation) **元素量**(1-N)			**和文件相關的註釋檔案資訊**

元素名稱					**定義**
			類型_*屬性*(type*)		**指出註釋的類型**(SEGMENT(**斷詞/句**): **包含斷詞或斷句的註解檔案**；GRAM (**文法**): **文件裡有關詞彙的構詞－句法資訊註解檔案**；ALIGN: **連接相對應翻譯的註解檔案**)
			註解.位置_*屬性*(ann.loc*)		**提供有關註解檔案的位置，例如路徑/檔案名稱、**URL **等**
			翻譯_*屬性*(trans.loc*)		**針對註解檔案裡包含了相關連的資訊**(alignment information)**，提供有關包含此相關連文件檔案的位置，包括路徑/檔案名稱、**URL **等**
改版描述(revisionDesc)					**摘要敘述一個檔案的改版歷史**
	變更(change) **元素量**(1-N)				**摘要描述一份多位研究者所有的電子文件的改版變更內容**
		變更日期(changeDate)			**給與變更的日期**
			日期表示_*屬性*(value*)		**以** ISO 8601 **的格式表示日期**
		變更者(respName)			**指出變更內容的人的名稱**
		h.**項目**(h.item)			**指出改變的性質，可以出現一或多個此類元素在每個<變更>元素內**

附件三

Spoken Dutch Corpus (Corpus Gesproken Nederland - CGN)

** http://www.mpi.nl/world/ISLE/overview/overview_frame.html **

Last update: 27-2-2001

- **簡介**

  荷蘭口語語料庫計畫主要目標在於建立荷蘭與法蘭得斯(Flanders)兩地當代成人荷蘭口語語料庫。待製作完成，語料庫將包含約一千萬字，2/3 來自荷蘭，1/3 來自法蘭得斯。荷蘭口語語料庫包含大量的錄音口語文件，大約錄有 1000 小時。

- **參考文獻**

  The Spoken Dutch Corpus (CGN-project)

- **後設資料概述**

元素名稱					定義
語料庫標目					包含有關計畫的一般資訊和/或適用於所有例子的資訊
	類型*				說明標目所屬的檔案類型
	建立者*				負責建立目錄的機構
	版次*				標目的版次
	更新日期*				最近更新標目的日期
	檔案描述(fileDesc)				?
		標題敘述(titleStmt)			有關語料庫內容的資訊
			標題		?
			負責人/單位(respStmt)		?
				工作類型(respType)	描述個人或單位負責的工作內容
				負責單位	負責機構的名稱
		版本(editionStmt)			發表的編號
			發表*		?
			版次*		?
		大小(extent)			語料庫大小
			字數(wordCount)		語料庫的所有字數
			秒數		語料庫的所有秒數
			位元數(byteCount)		組成語料庫的所有位元數
			大小補述(extNote)		某種計算方式的補充資訊，例如標點符號
		平均速度(tempoAV)			語料庫內說話的平均速度
			wph*		每一小時的平均字數
			(id)*		辨別組成成份的識別碼

元素名稱					定義
		出版單位 (publicationStmt)			有關語料庫出版或發行的資訊
			發行者/單位 (distributor)		發行者或單位的名稱
			發行者/單位地址 (pubAddress)		發行者/單位的地址
			電話(telephone)		發行者/單位的電話號碼
			傳真(fax)		發行者/單位的傳真號碼
			電子地址(eAeeress)		發行者/單位的 email 帳號
			可獲得性(availability)		語料庫真正版本的分布地區
				領域*	電子文件所採用的權限範圍
				層級*	提供文件現階段取得權限的識別碼
			出版日期(pubDate)		發行日期
			版權		版權所有者的名稱
	編碼描述 (encodingDesc)				記錄文件和來源的關係
		計畫描述 (projectDesc)			CGN 計畫的描述
		取樣陳述 (samplingDecl)			取樣方法的描述
		編輯陳述 (editorialDecl)			有關文件數位化和標記的資訊
			轉換(transduction)		描述數位化時轉寫錄音記錄的過程
			斷詞(segmentation)		描述語料庫內斷詞的原則，例如是以說話者、發音、句子、詞等
		關係陳述(refDecl)			描述斷句如何命名以及如何相關連
		類別描述(classDecl)			描述語料庫內樣本的分類
		元素量(1-N)	類型(category)		？
				類型描述 (catDesc)	？
	背景描述 (profileDesc)				有關語料庫的特殊資訊
		語言(langUsage)			描述語料庫內所包含的語言
		書寫系統 (wsdUsage)			包含一到多個<書寫系統>次欄位
		1..N	書寫系統		指出使用何種 ISO 字元
	改版描述(revDesc)				描述檔案的改變
	元素量(1-N)	變更(change)			？
			更新日期(Date)		變新的日期
			respName		？
				變更內容 (respType)	描述變更的日期
				resp	變更描述

元素名稱					定義
				負責機構名稱	負責機構的名稱
文件標目					？
	類型*				說明標目所屬的檔案類型
	建立者*				負責建立目錄的機構
	版次*				標目的版次
	更新日期*				最近更新標目的日期
	檔案描述(fileDesc)				包含語料庫書目資料的描述
		標題敘述(titleStmt)			有關文件(fragment)內容的資訊
			標題		？
			負責人/單位(respStmt)		？
				工作類型(respType)	描述個人或單位負責的工作內容
				負責單位	負責機構的名稱
		大小(extent)			文件大小
			字數(wordCount)		文件的所有字數
			秒數		文件的所有秒數
			位元數(byteCount)		組成文件的所有位元數
			大小補述(extNote)		某種計算方式的補充資訊，例如標點符號
		平均速度(tempoAV)			說話的平均速度
			wph*		每一小時的平均字數
		出版單位(publicationStmt)			有關文件出版或發行的資訊
			發行者/單位(distributor)		發行者或單位的名稱
			可獲得性(availability)		文件的傳布
				語料庫	？
				cd	？
				日期	？
		來源描述			有關來源的書目描述
			書目結構(biblStr)		書目資料描述
				作者	作者名字的起首字母與姓氏
				題目	題目
				出版社名稱	出版社名稱
				出版地	出版地點
				出版年代	印刷發行的年代
			rec		？
				日期*	錄音日期
				時間*	錄音時間
			來源		材料源出地
			製作者		製作錄音者
	編碼描述(encodingDesc)				記錄文件和來源的關係
		編輯陳述(editorialDecl)			在為文件編碼時的編輯原則與應用
			修正		
				類型*	？

元素名稱						定義
				現狀*		以 yes/no 回答
	背景描述 (profileDesc)					有關語料庫的特殊資訊
		文件類型(textClass)				指出文件相關的分類
			catRef			?
				目標*		一或多種對文件類型的描述
				關鍵詞*		自有限的列表裡挑選關鍵詞
				術語*		?
		參與者描述 (particDesc)				?
			說話者			?
				識別碼*		?
				角色*		說話者的角色
				年齡*		在錄音時說話者的年齡
			互動			參與者的互動
				類型*		
				主動*		主動(被識別出的)說話者的數目
				被動*		被動(未被識別出的)說話者數目
			關係			說話者間的關係
				主動*		在有直接關係時，對主動說話者的識別；或在非直接關係時的所有說話者
				描述*		關係的描述
				mutual*		indicates whether the relation holds fo all speakers or is directional
		背景描述(settDesc)				?
			地區			進行錄音的地區
			地點			進行錄音的地點
			場所(locale)			描述進行錄音的空間場所
			活動			簡短描述說話者所進行的活動
		錄音條件 (recCondition)				?
			錄音媒介(recMedium)			?
				類型*		錄音的媒介
				麥克風		用來錄音的麥克風類型
				錄音距離		?
					人	說話者識別號
					距離	?
					公分	距離
				噪音		描述錄音時的背景噪音
			數位化			?

元素名稱					定義
				opname	類比／數位？
				verwerking	類比／數位？
				現況	類比／數位？
	改版描述(revDesc)				描述檔案的改變
	元素量(1-N)	變更(change)			？
			更新日期(Date)		變新的日期
			respStmt		？
				變更內容(respType)	描述變更的日期
				resp	變更描述
				負責機構名稱	負責機構的名稱
參與者標目					？
	類型*				說明標目所屬的檔案類型
	建立者*				負責建立目錄的機構
	版次*				標目的版次
	更新日期*				最近更新標目的日期
	參與者描述(particDesc)				包含說話者一般資訊的描述
		人			？
			識別碼*		說話者識別碼
			性別*		說話者性別
		生日			？
			年*		說話者的出生年
			地點*		說話者出生地
			地區*		說話者出生的地區
		語言			？
			第一語言		說話者成長過程中所使用的語言
				語言*	？
				方言*	？
			家用語		說話者在家使用的語言
				語言*	？
				方言*	？
			工作語		說話者在工作時所使用的語言
				語言*	？
				方言*	？
			居住地		？
				地點*	說話者居住的地點
				地區*	說話者居住的地區
				人口多寡*	說話者居住區人口的多寡
			教育背景		？
				地點*	說話者受教育的地點
				地區*	說話者受教育的地區
				opleiding*	說話者的最高教育程度
				程度*	教育程度
			職業		？

元素名稱					定義
				工作*	說話者的職業
				層級*	職業的層級
			備註		其他有關說話者的補充說明，例如：參與的其他計畫、其他居住地等。

附件四

Metadata Elements for Session Description (v2.5)

⊙ 場次欄位(Session schema)名稱

藍字表示具有「次元素」

IMDI 欄位		
場次 Session		
名稱 Name		
標題 Title		
日期 Date		
地點 Location		
		洲 Continent
		國家 Country
		地區 Region
		地址 Address
描述 Description		
關鍵詞組 Keys		
計畫 Project		
	名稱 Name	
	標題 Title	
	識別號 ID	
	聯絡方式 Contact	
	描述 Description	
收集者 Collector		
	姓名 Name	
	聯絡方式 Contact	
	描述 Description	
內容 Content		
	溝通背景 Communication Context	
		互動 Interactivity
		預定類型 Planning Type
		介入 Involvement
	類型 Genre	
		互動的 Interactional
		無層次的 Discursive
		表達 Performance
	任務 Task	

	模式 Modalities	
	語言 Languages	
		描述 Description
		語言 Language
	描述 Description	
	關鍵詞 Keys	
參與者 Participants		
	描述 Description	
	參與者 Participant	
		類型 Type
		姓名 Name
		全名 Full name
		代號 Code
		角色 Role
		語言 Language
		種族 Ethnic Group
		年齡 Age
		性別 Sex
		教育程度 Education
		匿名 Anonymous
		描述 Description
		關鍵詞 Keys
資源 Resources		
	媒體檔案 Media File	
		資源連結 Resource Link
		大小 Size
		類型 Type
		格式 Format
		品質 Quality
		錄製狀況 Recording Conditions
		起訖位置(Position)
		版權 Access
		描述 Description
	註解單位 Annotation Unit	
		資源連結 Resource Link
		媒體資源連結 Media Resource Link
		註解者(Annotator)

		日期 Date
		類型 Type
		格式 Format
		內容編碼 Content Encoding
		字元編碼 Character Encoding
		版權 Access
		語言編號 Language ID
		匿名 Anonymous
		描述 Description
	來源 Source	
		編號 ID
		格式 Format
		品質 Quality
		起訖位置 Position
		版權 Access
		描述 Description
	匿名 Anonymous	
		資源連結 Resource Link
		版權 Access
參照 References		
	描述 Description	

⊙ **次元素(Sub-schemas)**

語言 Language	
	編號 ID
	名稱 Name
	描述 Description
關鍵詞組 Keys	
	關鍵詞 Key
關鍵詞 Key	
	名稱 Name = 值 Value
	字彙連結 Vocabulary Link
描述 Description	
	內文 Text
	語言編號 Language ID
	資訊連結 Info Link

版權 Access	
	可用性 Availability
	描述 Description
	日期 Date
	所有者 Owner
	出版者 Publisher
	連絡方式 Contact
連絡方式 Contact	
	姓名 Name
	住址 Address
	E-mail
	所屬機構 Organization