VIETNAMESE WORD SEGMENTATION

I. Definitions of Vietnamese words

1. *"A word is the smallest unit which has complete meanings and a fixed structure in order to form sentences."* [1]
2. "*A word is a combination of morphemes (or an entry) that appears in the Vietnamese dictionary*[1]*"*

II. Principles of annotating word boundaries

The general principle of annotating word boundaries is based on the two definitions above. However, we have to use the suitable one depending on the context given.

1. Basic words

It is based on the Vietnamese dictionary to define this kind of words (words that appear in the dictionary)

E.g.*: bàn [table], ghế [chair], sách [book], vở [notebook], học_tập [study], cơ_hội [opportunity], nhân_dân [people], công_nghiệp [industry], đại_học [university], kinh_tế [economy], xã_hội [society], …*

2. Idioms, locutions

- If idioms, locutions are entries in the dictionary, they are annotated as words.

  E.g.: *"rán_sành_ra_mỡ", "ăn_cây_nào_rào_cây_đó"*

- We just annotate word boundaries in higher levels, such as proverbs, folk-songs and sentences (simple sentences, complex and compound sentences).

Note: Apart from phrases which are real idioms (Vietnamese, Sino), the phrases in which the Sino factors dominate will be also clustered and annotated as an ID, such as: *Tứ_nhiếp_pháp/ID/O; Thiên_Long_Bát_Bộ/ID/O; Kiến_văn_tiểu_lục/ID/O*; …

3. Special cases

3.1 Classifiers and categorization words[2]

- Words indicating a single entity can be classifiers and they cannot be combined directly with numerals.

  E.g.: *máy [machine], hoa [flower], cá [fish], bệnh [sickness], pin [battery],…*

  We cannot use "*hai bệnh lao" [two tuberculosis],* but "*hai **loại** bệnh lao" [two **kinds** of tuberculosis]*

[1] The dictionary in use here is "Vietnamese dictionary – Professor Hoàng Phê, Science and Society Press, 1998"

[2] Refer to list of categorization words which are separable and từ chỉ đơn vị quy ước in the appendices.

- For these word combinations, we will consider the possibility of their short forms.

  E.g.:
  - *"bệnh tâm thần" [mental illness]* has the short form *"tâm thần"* (remove *"bệnh"*)
  - *"hoa hồng" [rose]* has the short form *"hồng"* (remove *"hoa"*)

- However, for machine categories: *máy may [sewing machine], máy in [printer], máy đào [digging machine], máy xay [grinder],…,* there are no short forms like *"may", "in", "đào", "xay"…*

- For those phrases which have short forms (high popularity and loose combination), we segment them into basic words.

  E.g.:
  - *bệnh tâm_thần[mental illness]*
  - *hoa[flower]/Nn hồng [rose]*

- Those phrases which do not have short forms will be considered words.

  E.g.:
  - *máy_bay [plane]*
  - *máy_in [printer]*
  - *máy_xay[grinder]*
  - *cá_chép[carp]*
  - *cá_rô[anabas]*

- For nouns which include inseparable categorization words such as *máy_xay [grinder], máy_in [printer], cá_chép [carp], cá_rô [anabas]*,… , they are only accepted as level 1 nominative nouns, i.e. words constructed by two basic components (*máy[machine] + xay[grind], máy[machine] + in[print], cá[fish] + chép, cá[fish] + rô,…*). For level two noun phrases, they have to be segmented into basic words.

  E.g.:
  - *máy_xay sinh_tố[liquidizer]*
  - *máy_in phun [ink-jet printer]*
  - *máy_in kim[matrix printer]*
  - *cá_chép hồng[carp]*
  - *cá_rô đồng [anabas]*

3.2 Synonyms

When examining the corpus, if there are words bearing meanings similar to those of the dictionary entries, those words will be added as new words.

3.3 Foreign words

- Words taken from foreign languages (including languages of the minorities) are adopted into Vietnamese language.

E.g.*: internet, email, polymer, design, vak_wei, drôk_tue, Bih_KJhô, …*

- When the names of substances, illnesses, animals and plants, programmes (in general), works of literature and arts, holidays, … are written in foreign languages (English, French, …): *hydro, natrium, Asian_Song_Festival, Aie_confiance_mon_amour, Woman's_Day, Christmas_ Day, …*, they are also annotated as FW.

3.4 Proper nouns

- Person name, location name, organization name…
  E.g.: *Smith, Mary, Tuấn, Hà_Nội, Windows*,…

- Names of the oceans will be annotated Nr: *Thái_Bình_Dương/Nr [Pacific Ocean], Ấn_Độ_Dương/Nr [Indian Ocean]*, …

- Chemical symbols: *CH_4, CO, HCl, Fe …will* be annotated Nr.

III. Examples

Việc ghép[combine] tế_bào[cell] gốc[origin] đã[did] giải_phóng[free] bệnh_nhân[patient] tiểu_đường[diabetes] type 1 khỏi[avoid] phải[must] tiêm[inject] insulin mỗi[every] ngày[day], theo[according] các nhà nghiên_cứu[research] tại[at] Đại_học[university] y_khoa[medical] Northwestern_University_Feinberg. (The transplant of stern cells has freed the patients with diabetes type 1 from being injected insulin every day, according to the researchers at Medical university of Northwestern_University_Feinberg.)
Đa_số[majority] bệnh_nhân[patient] tiểu_đường[diabetes] type 1 đã[did] có_thể[can] ngưng[stop] tiêm[inject] insulin trung_bình[average] sau[after] hai[two] năm[year] rưỡi[half] chữa_trị[treatment] với[with] các tế_bào[cell] gốc[origin] của[of] chính[own] họ[them]. (The majority of patients with diabetes type 1 have been able to stop being injected insulin on average after two and a half years of treatment with their own stern cells.)
Trước[before] đây việc ghép[transplant] cho[for] bệnh_nhân[patient] tiểu_đường[diabetes] type 1 gồm[include] việc lấy[take] mẫu[sample] từ[from] nhiều[many] người[person] hiến_tặng[donate] tế_bào[cell] tiểu_đảo[islet], thứ[thing] tiết[secrete] ra[out] insulin, và[and] tiêm[inject] cho[for] bệnh_nhân[patient]. (Previously, the transplant for patients with diabetes type 1 included sampling from many people donating islet cells which secrete insulin, and injecting the patients with insulin.)
Trong[in] cuộc nghiên_cứu[research] mới[just] được[be] thực_hiện[implement] tại[at] Mỹ[the U.S.] này[this], các tế_bào[cell] gốc[origin] được[be] lấy[take] từ[from] chính[own] máu[blood] của[of] bệnh_nhân[patient] và[and] được[be] can_thiệp[interfere] trong[in] phòng[room] thí_nghiệm[experiment] trước[before] khi[when] tiêm[inject] vào[into] lại[again] cho[for] họ[they], tránh[avoid]

vấn_đề[problem] tìm[find] người[person] hiến tặng[donate]. (In this newly implemented research in the U.S., the stern cells are taken directly from blood of the patients and interfered in the laboratory before injected back to them, thus avoiding the problem of finding the donors.)

Tế_bào[cell] gốc[orgin] dường_như[seem] "khởi_động[start] lại[again]" hệ[system] miễn_nhiễm[immune] của[of] bệnh_nhân[patient] không[not] tấn_công[attack] các tế_bào[cell] tiểu_đảo[islet], thứ[thing] sản_xuất[produce] insulin tự_nhiên[natural] trong[in] cơ_thể[body], theo[follow] đồng_tác_giả[co-author] nghiên_cứu[research] Richard_Burt thuộc[of] Đại_học[university] y_khoa[medical] Northwestern_University_Feinberg ở[at] Chicago. (Stern cells seemingly "restart" the immune system of the patients, not attacking the islet cells which produce natural insulin in the body, according to co-researcher Richard Burt of Medical University Northwestern_university-Feinberg in Chicago.)