



Deal all,

=> Sun Maosong, Kiyong, Gerhard, Nicoletta, Laurent, Sue Ellen

Please find some comments, that (I hope) will improve the document.

**POINT#1 about section-3**

I think the section-3 must focus on the terms that are important for the WordSegmentation document.

More precisely, I suggest to define in this standard, the following terms:

- segmentation
- lemmatization
- word segmentation
- word segmentation unit
- token
- tokenization
- text
- (raw / annotated) corpus

I suggest to borrow some definitions from **LMF (or MAF or SynAF, they are identical)** for:

- lexicon
- word
- lemma

I think that you don't need to define the other terms.

**Sun response:** for people no experience in WS, we need a well-defined concept system.

**Gil:** could you give me the latest WORD version pf LMF (also MAF,SynAF)? I'll read them carefully.

## POINT-2

Please respect the ISO rules for crafting definitions.

That means:

- a) never use an article as a first word
- b) use the singular, on the contrary of 3.1.25 for instance
- c) do not mix definition with other statements like guidelines, restrictions, examples like in 3.1.44. The definition part must only contain the definition. Use "Note" and "Example" for other matters.
- d) use an efficient style for the micro-structure of the definition like the classical genus-differentia pattern.

accept

## POINT#3

For terms that are i) too controversial or ii) leading to confusion, I suggest, simply to avoid to define and use them.

In principle, accept it. But in detail?

It is the choice of the editor of an ISO document to refuse or accept a term. So, I suggest you to make use of this right.

## POINT#4

If you want to keep the current definitions, I have some comments about them:

A) the distinction between lexicon and dictionary is not an accepted distinction. This distinction is rather controversial. For some linguists, a dictionary contains definitions and a lexicon does not have necessarily definitions. For some others, a lexicon is a term that is more general than a dictionary. For some others (closer to your interpretation), a lexicon is the result of an activity called lexicography, and a dictionary, the result of dictionary making<sup>1</sup>. From this point of view, lexicography (as a sub-field of lexicology) is both seen as a preparation and a scientific study for dictionary making.

By the way, I contest the distinction. Being lexicography or dictionary making, the list of entries is **always a matter of convention**. Even in France, where we have a dictionary that is published by the government (i.e. "le dictionnaire de l'Académie Française"), it is frequent to use and to talk about entries that are in or outside this dictionary. **I don't agree to say that a lexicon is an abstract entity and a dictionary a concrete entity**, as you do in Figure-1. It seems to me that this is an illusion.

---

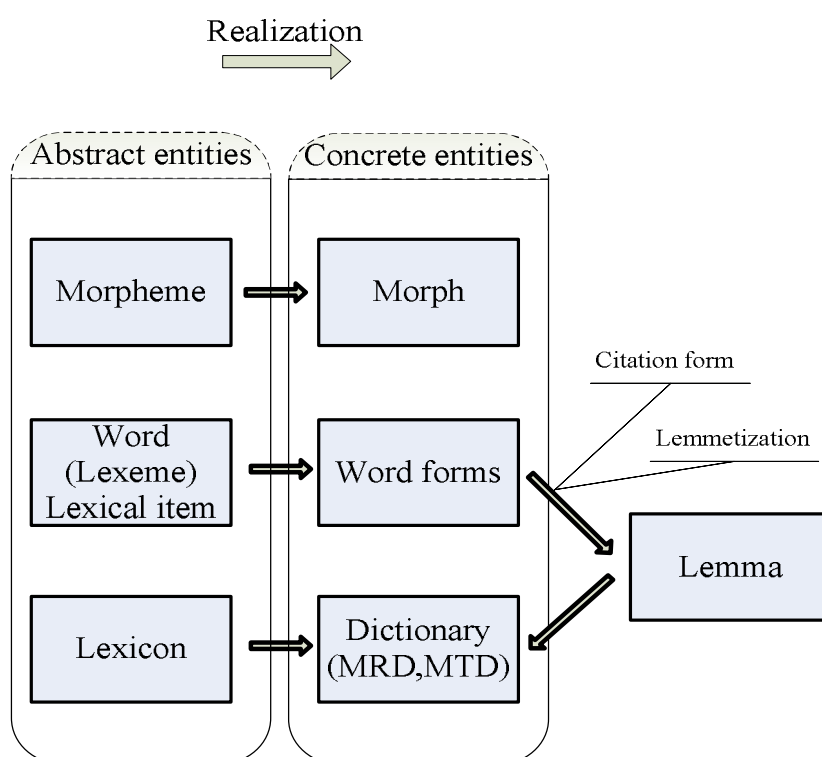
<sup>1</sup> Let's note that in French, we have a specific (and recently coined term by Bernard Quemada) for dictionary making that is "dictionnaire".

This is a fundamental issue which may affect the whole thing of this proposed standard. Face-to-face discussion is needed.

Which definition for lexicon in Wikipedia should be adopted?

Lexicon may refer to one of several things:

- A synonym for a dictionary or encyclopedic dictionary.
  - The lexical form or lemma (linguistics) of a word is its canonical form, under which it appears in dictionaries
- In linguistics, the lexicon is a language's inventory of lexemes.



B) The term "collocation" is, for other reasons very controversial. In the WordNet arena, a collocation is a compound because the Princeton team defined that in the foundation articles some years ago. But according to some other definitions, the scope is broader, like in LONGMAN-DCE: "the way in which some words are often used together". This is why in LMF, after some long talks, we decided to omit the term "collocation" and use the term "Multi-word expression" that is broader but not controversial. In fact, we don't need to classify the various types of MWE.

Tend to accept. Re-consider it .

C) In 3.1.21 "word formation", you forgot languages like the semitic ones where the formation is root based. In Hebrew or Arabic, for instance, the majority of words are built from a root and a scheme. **And the root is not a word.** It is not derivation because, **the source is not a word.** **And it's not inflection: it is clearly word creation.**

D) in 3.1.62, Lexicalization is defined as "the process of making a word to express a concept". Are you sure?

check

#### POINT#5

In the end of section "Scope": please, replace ISO WD 24613:2004 by ISO CD 24613:2006.

Also, replace ISO WD \*\*\* by ISO WD 24611

accept

#### POINT#6

In Figure-3, on the right side, you have "technical term" as opposed to "fixed expression". For me, it is not at the same level.

"technical term" is related to the creation and usage of a term, like "street term", "term for kids", or "everyday life term".

If you want to sub-class the term "multiword expression", you can use the three classes concerning the structure: "freely combination", "semi-fixed expression", and "fixed expression". It's of course more a continuum than a strict partition.

Accept, delete "technical term"

But the classification of multi-word expression cannot be done easily. What is possible to do is to define **dimensions of classification**. For instance: structure is a dimension. Having or not having a part of speech is another dimension. It is a little more complex than it appears to be at a first look.

#### POINT-7

Section 4.3: what is "word-hood" ? **delete it**

#### POINT-8

Concerning the definition of "stemming". You say that "it is usually sufficient that the related words" and after "Stemming can be regarded as an approximation of lemmatization".

The results are worst than that.

It's a well known fact that stemming is ONLY possible for simplistic morphology. May be the only inflectional language of this kind is English. Nobody use the stemming (as defined in your document) for other European languages like Italian, French or German, not talking about agglutinating languages. We know that lemmatisation without lexicon is very bad and we know that point at least since 30 years.

I don't see the rationale for the definition of "stemming".

Accept, delete "stemming"

#### **POINT-9**

Concerning the definition of lemmatization. You said that "in general lemmatization does not make sense for the isolating languages", but what's about recognition of multi-word expressions in Chinese?

How do you recognize that a certain sequence of strings is a multi-word expression if you don't do any lemmatization?

#### **POINT-10**

The principle in 4.1.1 of full coverage of text is circular.

The (word segmentation) standard should be applicable to any text that needs word segmentation.

Is it useful to state this circular statement?

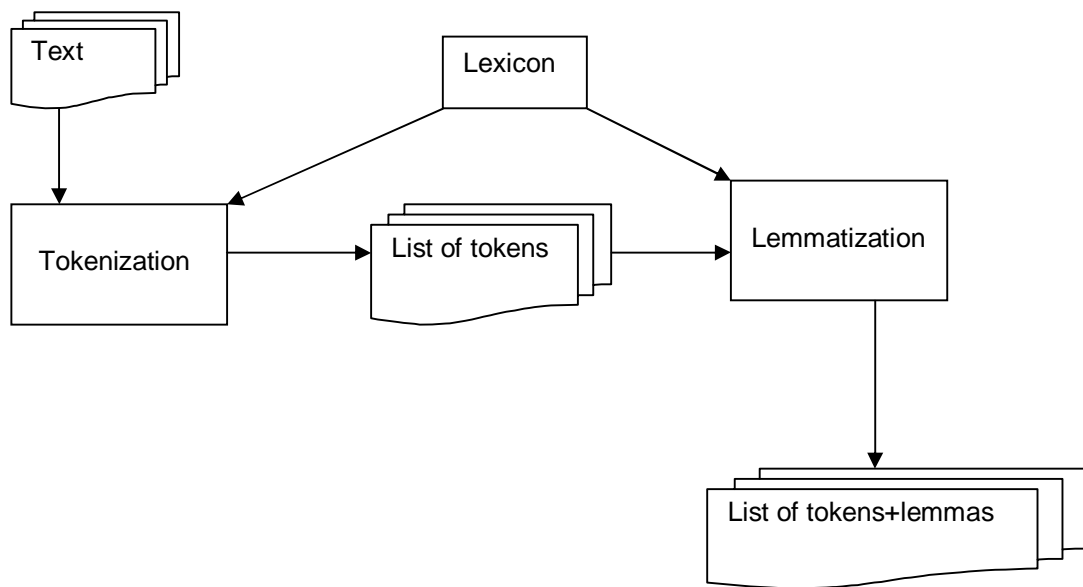
?

#### **POINT-11**

May be you can add a diagram like that in section 4 or 5?<sup>2</sup> Do you agree with this diagram (see next point) ?

---

<sup>2</sup> As you may know: I love diagrams ;-)



This is a very important point. Need to be discussed face-to-face.

#### POINT-12

There is a question that is not clearly stated and answered in section 4 & 5 of the document. **Is Word-segmentation the whole process of tokenization and then lemmatization? Or, is Word-segmentation only tokenization?**

**Basically, both will be done in the word segmentation process.**

Let's recall that lemmatization is defined as the process of determining the lemma for a given word form. **That means that multi-word expression recognition is to be done by the lemmatization. Lemmatization is not only for dropping affixes and fetching lemmas.**

In section 3.1.51, "word segmentation" is defined as "a process to generate an output with word boundaries annotated for any text". It's a little bit fuzzy. What is the output?

**Word forms (with structure)**

With regards to point-1, it is better to have a standard that has a small but good set of definitions rather than the contrary.

#### POINT-13

About 4.6.2, statement-2: "regular word forms should not be included in the lexicon, in general". This statement induces two comments. First, it's not in the scope of this standard to decide how a lexicon is internally managed. Secondly, from the strict point of view of lexicon use and management, if we split the words in three partitions: regular, exception and unknown ones. And, if we don't record regular forms, we obviously don't have any means to **DISTINGUISH** an exception from an unknown word(?). **This is why, usually (on the contrary of what is said) regular word forms are included in the lexicon.**

Lexicon is a collection of lemma, instead of all the word forms? Of course, word-forms of a lemma can be given in the lexicon under that lemma, but the level is different.

Is it the scope of this standard to have a section called: "general methods for word segmentation" ? From the discussion of this morning, I understood, that there is an agreement to drop the term "method" from the title, so the whole section 4.6 will be dropped. Isn't it?

Accept. Has changed to "basic concepts and general principles"

.....hope that helps, Gil