Constant-Sense Connection Paths

Einat H. Nir and Geoffrey L. Melnick

Geoffrey L. Melnick, Ltd., Patent Agents
17 Binyamin Mi-Todelo St., Tel Aviv, Israel
al_aluf@netvision.net.il

Abstract

A multilingual sense code may chart "constant-sense connection paths" across languages. A writer, not versed in any target language, may nonetheless proofread the sense for translation and edit it, to ensure that his meaning is conveyed as he wishes it, to other languages. A translation-ready format may be thus produced, to serve as a printing-press plate, for precise and automatic translation to any language, or to a plurality of languages. The translation-ready format may describe each word and the full document with a comprehensive code, which specifies the multilingual sense code and other relevant information about the word, in a standardized fashion, digitally, forming a unified, language-independent tagging system and a unified, language-independent lexicon.

Keywords

Word Sense Disambiguation, translingual, multilingual sense code, digital coding, comprehensive code, tagging, constant-sense connection paths, patents

Introduction

At times, we search for a specific word; there is a definite idea we wish to convey, but we miss the right word for it, so we consult with a dictionary, a thesaurus, and the like, until we find the exact word for the sense in mind. By what manner, code, or register, is the sense, not yet framed by a word, defined in the mind?

Word Sense Disambiguation (WSD) is based on the idea that words can be mapped into senses, Miller (1993), Dolan (1994), Dras (1997), Banerjee (2001). At its incipient stage it seemed promising.

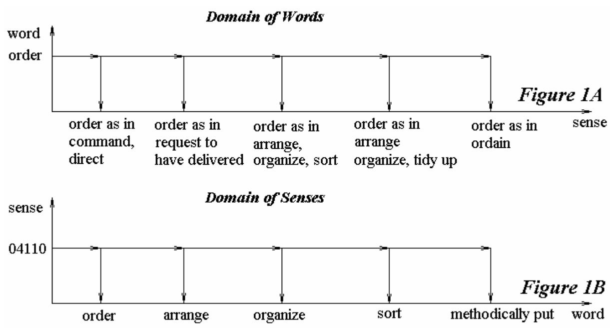
Yet Kilgarriff & Rosenzweig (2000) and Kilgarriff & Tugwell, (2001) report that the best performance of automatic systems, for English, in terms of WSD, is about 77%, compared with over 95%, for humans. They further suggest that a plateau has been reached, regarding WSD. Furthermore, Kilgarriff (1997), argues that word senses "are known to be very slippery entities," and "may only exist relative to a task." On Internet forum, Kilgarriff (2005) writes, "The trouble with Word Sense Disambiguation is word senses. They just won't behave."

Writers in various fields spend endless hours editing and proofreading a text. Yet, the translated versions, by humans or by machines, may convey a very different meaning from that which they intended, and when not versed in the language of the translation, the writers may be entirely unaware of that fact.

The Domain of Words and the Domain of Senses

In engineering, transforms between domains, such as space and time or polar and Cartesian coordinates are routine. They switch between the independent and the dependent variables and provide one with a different point of view.

When we communicate in words, words are the independent variables, and their multiple senses are the dependent variable. For example, for the word "order" as an independent variable, the senses, "order as in command," "order as in request," and others, are the dependent variables. (See Figure 1A.) When the word "order" is to be translated, an algorithm will be required to sort out the proper sense.



In contrast, if we communicated in senses, senses would be the independent variables, and the words associated with them, the dependent variables. (See Figure 1B.) For example, a sense "04110," which for the sake of the present example means "to put into a methodical and systematic arrangement," is associated with several synonyms, such as, "order," "arrange," "organize," "sort," "methodically put," and possibly others.

When the sense 04110 is to be converted to a natural language, another algorithm will be required, to choose from amongst the different synonyms. While a particular choice may influence style or language register, no ambiguity should arise.

When one operates in the domain of senses (see Figure 2A), converting from a sense in a source language to a sense in a target language, one is moving along a "constant-sense connection path," and there is one-to-one correspondence between the source and the target senses.

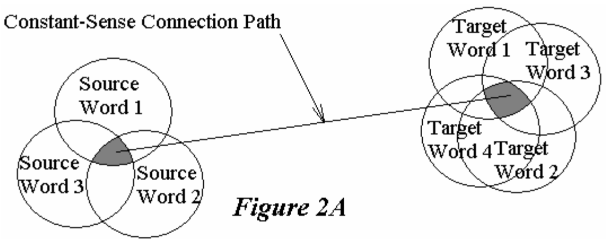


Figure 2A

However, when one converts from a sense in a source language to words in a target language (see Figure 2B), there is one-to-many correspondence between the source and the target, since the sense may be associated with several synonyms.

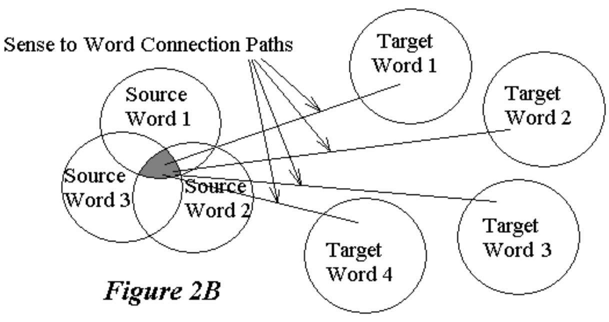


Figure 2B

Indeed, target words 1, 2, 3, and 4 may all be synonyms. But the operating utility does not know that. It sees a divergent situation, no different from that arising in word-to-word paths. This may be the reason word senses seem "slippery entities," that "just won't behave."

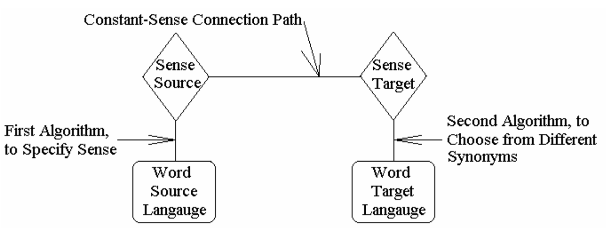


Figure 2C

A translation path should be as illustrated in Figure 2C:

1. from a word in a source language to a sense, as defined by the source language, via a first algorithm, for specifying the sense;
2. from the sense, as defined by the source language, to a sense, as defined by a target language, along a constant-sense connection path; and
3. from the sense, as defined by the target language, to a word in the target language, via a second algorithm, for selecting the target word from amongst different synonyms.

Constructing a Sense Code for Constant-Sense Connection Paths

Figure 3 illustrates a possible mapping of English words into senses and their grouping into synonym clusters. Each synonym cluster is assigned a sense code and is associated with synonym clusters in other languages. For example, the word "order" may be clustered with "arrange" and "organize," assigned a sense code, such as 04100 and linked to synonym clusters in Hebrew, and German, of substantially the same sense. Naturally, other languages may be included as well.

Other monolingual definitions may also be used, in place of the synonym clusters, for example, "to order - to put into a methodical arrangement."

Figure 3 is a representation in the domain of words.

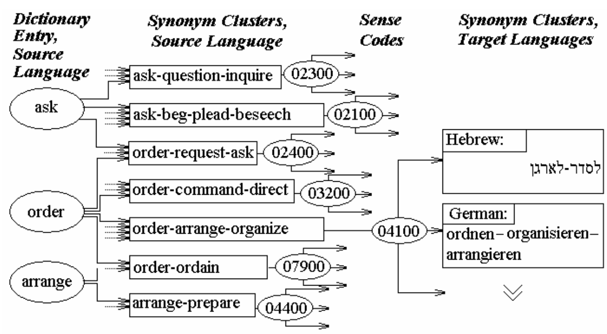


Figure 3

A transform to the domain of senses will result in the representation of Figure 4, arranged by sense codes. Figure 4 provides digital sense codes and their associated synonym clusters in different languages, for example, English, Hebrew, and German and represents a multilingual sense dictionary, which may serve as a basis for constant-sense connection paths.

04100 English: order-organize-arrange
Hebrew: לסדר - לארגן
German: Ordnen-organisieren-arrangierren

04110 English: order-organize-arrange - sort
Hebrew: לסדר - לארגן - למיין
German: ordnen-organisieren-arrangieren- sortieren

04120 English: order-organize-arrange-tidy up
Hebrew: לסדר-לארגן-לנקות

Figure 4

Naturally, some languages may describe more senses than others. For example, cold countries generally have several terms for snow, which are absent in languages of warm countries. In such cases, definitions may be used. For example, a sense 08900 may represent "snow," in both English and Hebrew, while 08910 may represent "slush" in English, and the equivalent of "partly melted or watery snow," in Hebrew, since Hebrew does not have a word for snow as "slush."

A constant-sense connection path may also be represented as a multilingual sense ring of synonym clusters or monolingual definitions, which share a common sense, in different languages, as seen in Figure 5A.

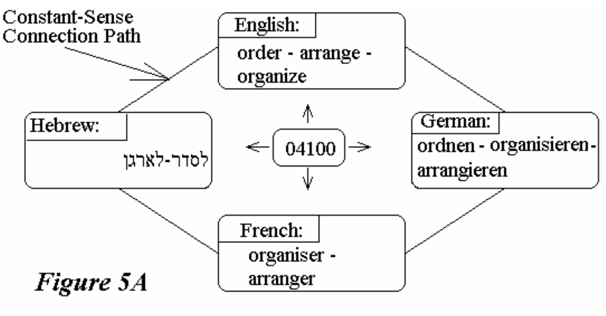


Figure 5A

A somewhat less desirable situation is illustrated in Figure 5B, where a monolingual sense definition is provided only for a single language, yet based on that sense definition, connection paths are charted to other languages, each defining a single word or term, in the other languages. As such, the divergent situation of Figure 2B may be avoided. But writers in the other languages will not be able to proofread and edit the sense for translation.

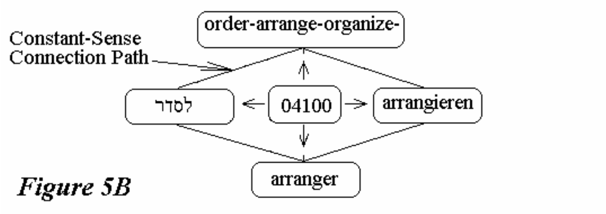


Figure 5B

Idioms and other terms should be treated as words, or dictionary entries, although in general, an idiom tends to have a fixed sense. For example, the idiom "sleep on it" may be defined by another idiom, "think it over" and (or) by the words "consider," and "reflect."

Strictly speaking, Figures 5A and 5B do not represent attempts to define senses in terms of multilingual correspondences, as suggested, for example, by Resnik & Yarowsky (1999), Ide et al.(2002), Dyvik (2004), Diab (2003), and others, for example, by aligning parallel corpora. Rather, in Figures 5A and 5B, the senses have been defined, in a monolingual manner, in the source langauge, to enable proofreading of the sense choices in the source language. Cross-linguistic correspondence to senses in other languages is made after their definition, to provide the constant-sense connection paths. Nonetheless, a synergetic process between the monolingual and multilingual definitions may take place.

The manner of manually sense coding a document, or proofreading a sense-coded document, in soft format, using the multilingual sense-code dictionary, is illustrated in Figure 6. The writer is presented with the several sense choices and confirms or changes the selected choice.

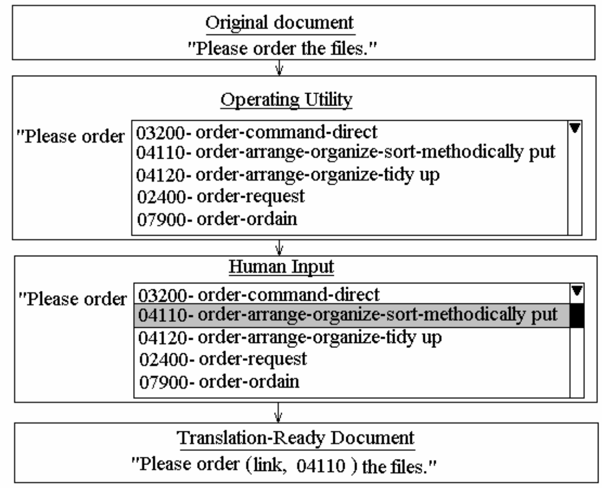


Figure 6

Thus, a writer, versed only in the source language, may nonetheless proofread and edit the senses for translations, and ensure that the translation will communicate the meaning that he intended.

At the Target-Language End

Sense coding may identify a synonym cluster in the target language. (See Figure 7A.) For a transform back to the domain of words, in the target language, a new algorithm is needed, for choosing a word from amongst the different synonyms, which convey the same sense.

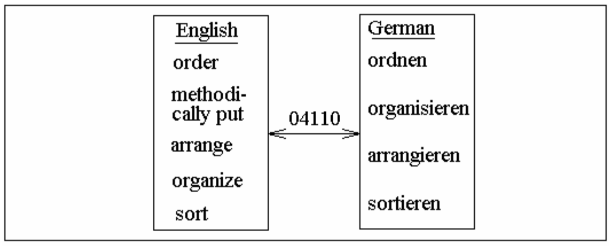


Figure 7A

The particular choice may influence style and language register. There are several options for choosing from amongst the synonyms of the target language. Where there is a common stem, for example, as in “organize” and “organisieren,” that may be used for association. (See Figure 7B.)

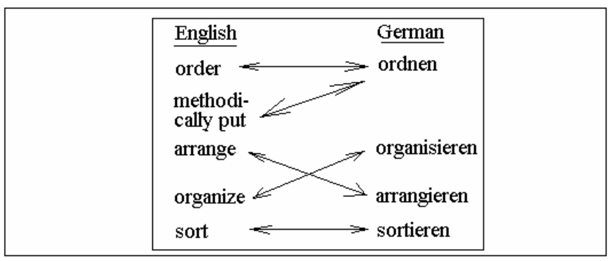


Figure 7B

Where the synonyms in the cluster are essentially equivalent, a randomizer may be used, to choose from amongst them. Yet, the randomizer may be instructed to remain with a particular choice for a given document. Where the synonyms are of a different register, translation may proceed at a specified register level, (see Figure 8)

Low Register	*Medium Register*	*High Register*
common - 85 %	common - 85 %	common - 84 %
widespread - 15 %	widespread - 10%	widespread - 10 %
	prevalent - 5 %	prevalent - 4 %
		ubiquitous - 2 %

Figure 8

using a skewed randomizer, to reflect the frequency of use of the different synonyms in the natural language. For example, a randomizer, operating at a medium register in English, may be instructed to use "common" 85% of the times, "widespread" 10% of the times, and "prevalent" 5% of the times.

Given the multilingual sense code of Figure 4, the manner of choosing amongst synonyms is the final step in the translation process, leading to a single word in the target language, which is substantially of the same sense as that of the original word in the source language.

Words and Senses the Fluid and the Discrete

If polysemous words were collections of discrete senses, mapping each for the associated senses would be a manageable task. However often words seem to be fluid entities, with a spread of senses about them, each sense spilling over and mingling with others.

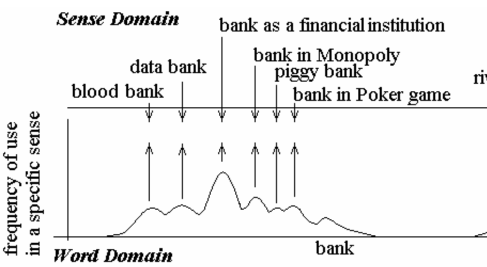


Figure 9

For example, the word "bank" (Figure 9) has a "sense spread" about it, relating to holding reserves, another, relating to natural slopes, and other senses as well. Mapping all the discrete senses associated with "bank" is probably, an impossible task, and there are likely to be senses that will be overlooked.

We wish to suggest, somewhat boldly, that senses exist, as discrete entities in our minds, as evident by the fact that we sometimes have a clear idea of a sense we wish to express, but search for the right word for it. Words are the tools we use to reach our sense registers. Words are fluid, and have a spread of meanings about them, because our minds work by association, and it is the very spread in meaning that provides the association. Had our minds worked like computers, words would have been precise digital units.

Humans have invented words to express senses, which seem to be coded in their minds by multiple associations. The nature of the relationship between words and senses is such that humans can always put senses into words. But the opposite is not true. Expressing words in terms of all their senses may not always be possible.

The multilingual sense dictionary (Figure 4) brings senses out into the open, and codes them in a language-independent, digital manner, so as to convert the "multiple-association" register of the mind with an unequivocal code.

In the multilingual sense dictionary, ***each entry has a single meaning***, making it a powerful, multilingual communication tool.

Unlike the InterLingual Index (ILI) of Euro Wordnet, which includes equivalent relations between synsets in different languages, within Euro Wordnet and Princeton's WordNet, and which has been used for information retrieval, the multilingual sense code, proposed here, is used directly for sense tagging, to provide the constant-sense connection paths across languages.

There is another important difference with the ILI. According to Gilarranz, et al. (1997), two approaches had been considered in consolidating Euro Wordnet and WordNet. The first, to introduce new synsets in WordNet structure (such as WordNet 1.5), where a sense in one of the European languages does not have a corresponding sense in WordNet. In such a case, every sense, in any one of the languages of Euro Wordnet and WordNet, would be linked to corresponding senses in the other languages.

The second was not to augment WordNet, and express the additional senses by complex cross-language translation links.

The second option was preferred, to avoid the need for a consensus in introducing new senses to WordNet. In consequence, the ILI is not a true interlingual dictionary, as suggested by Figure 4, and connection paths between languages are more difficult to establish.

In other words, issues such as politics and copyright, can stand in the way of producing an effective network of connection paths across languages.

Exposure-Based Learning in Real Time

Early on, lexical information was based on dictionaries. But as Atkins (1991) points out, machine-readable dictionaries (MRD) include human bias, which makes word mapping very difficult and of limited value. The definitions may be ambiguous, and there are inconsistencies among the different dictionaries, with respect to the mapping of words into senses.

Another strategy for obtaining lexical information is extracting information from corpora. This strategy has been growing over the past ten years, and as Kilgarriff (2003) points out, a by-product is the development of technologies for finding the lexical facts that go in dictionaries.

In parallel corpora, the same texts are provided in two or more languages, and source and target language words can be linked. This has been suggested, for example, by Resnik & Yarowsky (1999), Ide et al.(2002), Dyvik (2004), and others.

We wish to propose still a fourth strategy for lexicon acquisition - learning by exposure, in real time, relying on text, which at present must undergo manual translation. In essence we suggest the following:

- providing a text in a source langauge, for translation;
- sense coding the text;
- proofreading and editing the sense coding, by the writer, or by a person connected with the work;
- translating the text to one or several target languages;
- linking the source words, the proofread sense codes with translation words, to form sense code rings, for example, as illustrated in Figures 5A or 5B; and
- cataloging and linking the source and translated documents, in soft format, for future reference.

The suggested method will enable a writer to proofread his work, and at the same time, contribute lexical information of proofread senses, in context, or "***relative to a task***."

The corpus we suggest is patents and patent applications, which cover a wide range of fields, such as toys, engines, communication devices, medical implements, drugs, and genes, and which at present undergo manual translations.

Consolidation of the lexical information acquired from this corpus with WordNet, Euro Wordnet and others will be attempted.

Comprehensive Code

A thrifty, integrated tagging tool is proposed - a comprehensive code, formed as a vector of "n" natural numbers A(1), A(2)….. A(n). An example is described in Table 1. Each natural number "n" is expansible, as necessary, represented as a single digit, a double digit, and so on. Alternatively, real numbers may be used.

The comprehensive code will specify in a standardized fashion, digitally, substantially all the relevant information in regard to a specific word and its relations with other words in a sentence, for machine translation to any natural language. The information may relate to conjugation, pronoun references, and cultural matters, for producing a polished translation-ready format.

Table 1

A(n)	n	v	adj	adv	prep	conj	art
A(1)	sense code (e.g., 04110)						
A(2)	function						
	1	2	3	4	5	6	7
A(3)	$1=1^{st}$ person; $2=2^{nd}$ person; $3=3^{rd}$ person						
A(4)	1=male, 10=males, 2=female, 20=females, 3=object, 30=objects						
A(5)	translate phonetically						
A(6)	1= archaic, 2=colloquial, 3=legal, 4-unusual…						
A(7)	1= sub	1= pred	$1=1^{st}$, $2=2^{nd}$, adj	$1=1^{st}$, $2=2^{nd}$, adv.	$1=1^{st}$, $2=2^{nd}$, prep. phrs		1= a 2= the 3= said
A(8)	type *	tense	poss-ssive				
A(9)	1= name 2= prn,	1=act 2=pas 3=ref 4=int	1=act 2=pas				
A(10)	relating to a patent element						
A(11)	links words of a noun phrase by an index, if desired						
A(12)	links words of a verbal phrase by an index, if desired						
A(13)	links a preposition phrase by an index, if desired						
A(14)	links an idiomatic expression by an index, if desired						
A(15)	links a clause by an index, if desired						
A(16)	insert a translator's comment, if desired						

* 1=human, 2=animal, 3=plant, 4=object, 5=abstract, 6=action, 7=place,8=time,..

Some points need to be made, in connection with Table 1:

- Tenses are often language dependent, and the source language may need to be specified. Thus, the first tense digit may code the language, for example: 11=Eng, past simple, 111=Eng, past perfect, 1101=Eng, past continuous, 1111= Eng, past perfect continuous, 12= Eng, present simple, 121= Eng, present perfect, 1201= Eng, present continuous, 1211= Eng, present perfect continuous.
- Verbs and nouns often go with certain prepositions, e.g., "look at," "angry with," "mad at," "result in," "the result of." Treating them as verbal or noun phrases will ensure that the preposition association will be maintained.
- a digital phonetic code, which includes all the sounds and vowels of all the languages is needed, for phonetic translations, such as of names. New sounds need to be introduced to each langauge to cover these, in a systematic manner (e.g., kh in English for the Spanish "J"). When A(5) includes the digital phonetic code, the word will be translated phonetically, using the target-langauge alphabet.

In a sense, the comprehensive code will operate as a checklist, for verifying that all information relevant to the decision making of the operating utility will be available. In general, the comprehensive code may be formed automatically, by the operating utility. But where information is lacking, human input will be sought.

In terms of the bits, the comprehensive code is thrifty, using digits rather than words, which must be converted to ASCII and incorporating all tags into a single vector. The links, A(11) – A(15) overcome issues of order within a sentence, e.g., adjective-noun, versus noun-adjective.

The comprehensive code may be used as a multilingual tool for information retrieval and data extraction, as well.

Most important, because the comprehensive code is digitized, a single tagging system is applicable to all languages, and readable by all languages, so translation-ready formats between any two languages and for all languages may be produced, in a systematic manner.

Digital Morphology

The comprehensive code may serve as a basis for "digital morphology" using the sense code as a stem. Consider the examples:

Sam was a ***sorter***. He ***sorted*** files for a living. But the ***sorted*** files got to him. He was tired of ***sorting***. So he bought a ***sorting*** machine ***to sort*** his files. And he promised himself that he ***would never sort*** files again.

The comprehensive code can assign each of the underlined words the sense 04110 but different other attributes, for example:

For ***sorter***, in, " Sam was a ***sorter***":

- A(1) denotes the sense code, 04110;
- A(2) denotes a noun;
- A(3) denotes 3^{rd} person;
- A(4) denotes male;
- A(8) denotes a human;

sorter = 04110,1,3,1,0,0,0,1,0,0,0,0,0,0,0,0.

For ***sorted,*** in, "He ***sorted*** files for a living":

- A(1) denotes the sense code, 04110;
- A(2) denotes a verb;
- A(7) denotes predicate;
- A(9) denotes: active; and
- A(8) denotes past tense, simple.

sorted = 04110,2,3,1,0,0,1,1,1,0,0,0,0,0,0,0.

For ***sorted***, in, "But the ***sorted*** files got to him";

- A(1) denotes the sense code, 04110;
- A(2) denotes adjective;
- A(7) denotes first adjective;
- A(9) denotes adjective in a passive form.

sorted = 04110,3,3,1,0,0,1,0,2,0,0,0,0,0,0,0.

Although "sorted," as an adjective, does not appear in most English dictionaries, the comprehensive code enables one to generate a meaning for "sorted" as an adjective, by combining the in-context sense code of A(1) with the in-context function of A(2). This applies also to attributive nouns, for example, "care" in "health care costs, or "blood" in "blood flow rate studies." Digital morphology makes it possible to create words where necessary, by assigning them the in-context sense code, the in-context function, and other information of the comprehensive code, as relevant.

Digital Language ?

Apart from its use for tagging, the comprehensive code may be regarded as an expressive and computationally thrifty way of representing lexicons. It is generative, easily defining new senses and new forms, such as noun, verb, and the like, and it provides multilingual, unified definitions, which are unequivocal, each being singular in sense and form. As such it may be of relevance in discussions of The Generative Lexicon, Pustejovsky (1995), Bouillon & Busa, (eds.) (2001).

For example, the sentence, "He sorted files," can be expressed digitally, in a multilingual manner. "He" needs no sense code, being defined by the other attributes. Let the sense code for "file - a collection of papers arranged in a folder," be 06750.

Thus,
He sorted files. = 0,1,3,1,0,0,1,1,2,0,0,0,0,0,0
04110,2,3,1,0,0,1,1,1,0,0,0,0,0,0,0
06750,1,3,30,0,0,0,4,0,0,0,0,0,0,0,0.

The digitized sentence will be understood by any person, having a multilingual sense dictionary, such as Figure 4, which is inherently unequivocal, and a table of the vector notations, such as Table 1, where Table 1 is provided in that person's tongue.

The digitized sentence is our translation-ready format, to be converted to any natural language, on demand.

Translation-ready Formats

Unsupervised machine translation may have reached a plateau, in terms of accuracy, yet there are advantages to human-aided machine translation that have not yet been fully explored. These include the following:

1. creating a human-coded or human-proofread, translation-ready document that may serve as a printing-press plate, for translations to many languages, precisely and automatically, for example, for patents, scientific papers, periodicals, and the like;
2. creating a human-coded or human-proofread, translation-ready document for translation to any language on demand, for example, of a web site page;
3. enabling communication, for example, by e-mail and other soft formats, between people who do not

share a common tongue, by exchanging translation-ready documents;

4. providing writers with tools for controlling how their message is to be conveyed in other languages.

Figures 10A and 10B illustrate our visions for preparing translation-ready formats and for translations, in conjunction with a word processor.

Cataloging of the source and translated documents is desirable, for future comparison and studies. General document tagging, manually or semi-automatically, may define properties, such as, the source language, register level, professional field, and type, for example, third-grade reading material, legal matter, scientific paper, patent. After checking for audio-to-text input, grammar, and spelling, the document may be tagged for function, parsed, and sense coded. These may be performed manually or semi-automatically, with the operating utility seeking human input, as needed. Last, a human operator may review, accept, reject, or edit the translation-ready format, while working in his native tongue.

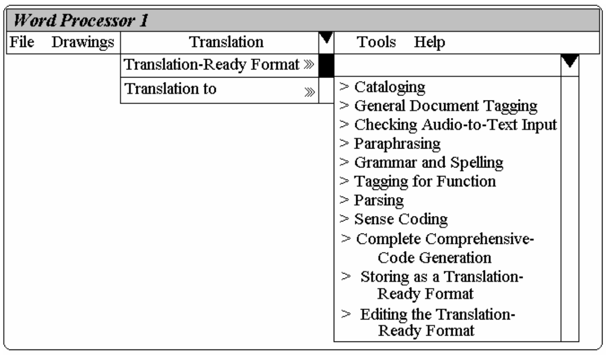


Figure 10A

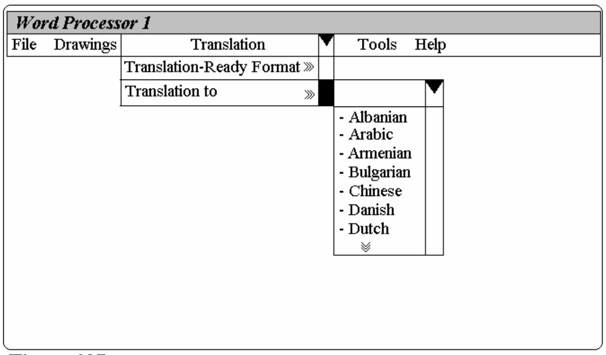


Figure 10B

Patents and Patent Applications a Test Case

Patent applications and the patents that mature from them are special documents, where subtlety is important. They undergo extensive prosecution, in which long arguments persist over how meaningful and significant the various elements and processes are, when compared with existing ones – the prior art, and they serve as bases for litigations, over intellectual property rights. By their nature, they need to be translated into many languages, and often, examiners and judges see only the translated version, while the inventor and his patent practitioner work from the original-language document. Generally, there is no communication between the inventor and his patent practitioner on the one hand, and the translator, sitting in the country where the translation is made, on the other. While the inventor and his patent practitioner may spend days, even weeks, on the exact formulation of their ideas, the translator works at a fixed fee per page, as fast as he can. The inventor and his patent practitioner can only hope that the examiner or judge sees in the translated version what they intended.

Accurate translation may be a deciding factor between allowance and rejection, and while in general, manual proofreading of tagging would seem time consuming, verification that a patent application or patent has been translated as intended, is time well spent.

The method we propose, of proofreading and editing the translation-ready format, will provide a writer with the tools to ensure that his ideas will be translated correctly.

At the same time, this method will build up an inventory of proofread sense codes, in context, and parallel translations to other languages, generally, from English to Japanese, Korean, Chinese, German, Italian and French.

Patents have a specific structure, where different portions employ different styles. The background and disclosure sections are written in scientific style, but in the disclosure, each element is assigned a number, shown also in drawings, and any reference to the element must be made with that number. The claim section, which defines the claims the patent makes to a specific intellectual property, is written in legal style. The various elements described in the disclosure must appear in the claims using the exact noun phrases, in which they appeared in the disclosure. In the US, they are listed without their assigned numbers, but in Europe, with the numbers.

We suggest that general tagging be used to indicate each of the patent sections, such as "Patent Background," "Patent Disclosure," "Claim set Europe," "Claim set US."

Specific words are important. Both "comprise" and "include" are used to describe incomplete listings, which can “further include” additional features, while “consist” refers to a closed group, that cannot be expanded. (This, at odds with The American Heritage® Dictionary of the English Language, which defines “comprise” as “consist”, stating that “*Comprise* usually implies that all of the components are stated.”)

In Europe, according to Rule 29[1](b) of the Implementing Regulations, the claim must include a “characterizing” section, using “characterized by” or “characterized in that” to separate between the known combination and the novel.

We recommend assigning specific sense codes to each of "characterized by," "characterized in that," "comprise," "include," and "consist."

As indicated, the various elements described in the disclosure must appear in the claims using the exact noun phrases, in which they appeared in the disclosure. Consider, for example, the following bit of disclosure: "an electrical apparatus 20, which includes a first proximal device 22 and a second proximal device 24, both receiving power from a powering mechanism 26, located at a distal portion of the electrical apparatus 20."

Strictly speaking apparatus, device, and mechanism are synonyms, yet here, each noun phrase has to be distinct. The translating utility should not introduce an electrical ***apparatus 20*** and later refer to the electrical ***device 20***.

The comprehensive vector may be used, to keep the patent elements distinct and consistent, for example, using A(10) to denote a patent element, and A(11) to link all elements of each noun phrase. In this way, references to each element in the patent and patent application will be distinctly maintained throughout.

Conclusions

A transform from the domain of words to the domain of senses is suggested, for the construction of a multilingual sense dictionary, in which senses are digitally coded. Each sense has a ***single*** meaning. When operating in the domain of senses, there are no ambiguities.

The digital sense codes may be used for sense tagging, or sense coding, in any langauge, and are operative as constant-sense connection paths across languages.

A lexicon-acquisition strategy - learning by exposure, in real time, is proposed, relying on text, which at present must undergo manual translation. Exposure-based construction of the multilingual sense code should provide valid and true-to-the-task word-sense resolutions.

A comprehensive code is presented, formed as a vector of n natural numbers, for specifying, in a standardized fashion, digitally, substantially all the relevant information of a word, for machine translation to other natural languages. The information includes the digital sense code, the function, conjugation information, pronoun-reference information, cultural matters, and special requirements, such as “translate phonetically.”

The comprehensive code is an expressive, computationally thrifty representation of a lexicon. It is expansible, or generative, easily defining new senses and forms, and it provides substantially unequivocal definitions, unified for all languages.

Because the comprehensive code is digitized, it is applicable to all languages, and readable by all languages; translation-ready formats between any two languages and for all languages may be produced, in a systematic manner.

The translation-ready format may be proofread and edited by a writer, giving him control over his expressions, in any language.

References

Miller et al. (1993) ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.pdf.

Dolan (1994) Linguistics, ACL, Morristown, N.J., pp. 712-716.

Dras (1997) ACL-EACL'97, 516--518. Madrid, Spain.

Kilgarriff & Rosenzweig (2000) Computers and the Humanities, 34(1-2): 15-48.

Kilgarriff & Tugwell (2001), MT Summit VIII, Santiago de Composter, Spain.

Kilgarriff (1997) url = "citeseer.ist.psu.edu/kilgarriff97dont.html."

Kilgarriff (2005) http://torvald.aksis.uib.no/corpora/2005-1/0455.html.

Atkins & Levin (1988) Proceedings of the Fourth Annual Conference of the UW Centre for the New OED, Oxford.

Atkins, (1991) International Journal of Lexicography 3:167-204.

Philip Resnik and David Yarowsky (1999) *Natural Language Engineer-ing*, 1(1):1–25.

Dyvik (2004) http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/ICAMEpaper.pdf

Ide, et al. (2002) Proceedings of ACL'02, Philadelphia, 54-60.

Diab (2003), PhD. thesis, U of Maryland.

Pierrette Bouillon and Federica Busa, (eds.) (2001), *The Language of Word Meaning*. Cambridge: Cambridge University Press, xvi, 387 pp.

Pustejovsky (1995) The Generative Lexicon ISBN 0-262-16158-3

Kilgarriff (2003) Proceedings of ASIALEX, ITRI-03-16.